UNIVERSITÉ DU QUÉBEC À MONTRÉAL

A DIRICHLET MIXTURE MODEL-BASED APPROACH FOR IDENTIFYING SPAMMERS IN ONLINE SOCIAL NETWORKS

MASTER THESIS
PRESENTED
AS A PARTIAL REQUIREMENT
FOR THE MASTER IN COMPUTER SCIENCE

BY
FARNOOSH FATHALIANI

MAY 2015

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UNE APPROCHE PROBABILISTE BASÉE SUR LA DISTRIBUTION DE DIRICHLET POUR L'IDENTIFICATION AUTOMATIQUE DES SPAMMEURS DANS LES RÉSEAUX SOCIAUX

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR
FARNOOSH FATHALIANI

MAI 2015

ACKNOWLEDGMENTS

It gives me great pleasure in expressing my gratitude to all those people who have supported me and had their contributions in making this thesis possible.

First and foremost, I would like to express my profound sense of reverence to my advisor Professor Mohamed Bouguessa, for his constant guidance, support, motivation, inspiration, enthusiasm, and immense knowledge. I could not have imagined having a better advisor and mentor for my master's study.

I am indebted to my friends and colleagues for providing a stimulating environment in which I could learn and grow, especially I thank my friends, Ali Sefidpour, Sima, Maryam, Mona, and Shaghyegh.

I would like to thank all the staff members of the Computer Science department at UQAM for their direct and indirect helps during my studies at UQAM.

Last but not least, I would like to thank my family: my parents, Khosro and Roshanak, for their love, encouragement, advice, and for supporting me spiritually throughout my life.

RÉSUMÉ

Ce mémoire propose d'étudier la problématique de l'identification automatique des utilisateurs malicieux (les spammeurs) dans les réseaux sociaux. Notre contribution consiste à développer un modéle probabiliste qui exploite le modéle de mélange de la distribution de Dirichlet pour détecter les spammeurs. Spécifiquement, dans notre méthode nous proposons d'estimer un vecteur de caractéristique pour chaque utilisateur d'un réseau social. En partant du fait que les spammeurs sont des utilisateurs avec des caractéristiques atypiques comparativement aux utilisateurs normaux, chaque valeur de ce vecteur relate ce que nous appelons le "degré d'anormalité" de chaque utilisateur, et ce, selon les différents modes d'interaction dans un réseau social. Les spammeurs devront avoir des valeurs de degrés d'anormalité trés élevées comparativement aux utilisateurs normaux. Pour discriminer les spammeurs des utilisateurs légitimes, nous proposons un modéle probabiliste qui s'appui sur l'utilisation des mélanges de distribution de Dirichlet pour estimer la fonction de densité de probabilité des vecteurs de caractéristiques. Le choix de la distribution de Dirichlet est principalement motivé par la grande capacité de cette distribution à modéliser des situations complexes et variées.

L'approche proposée posséde quatre mérites : (1) ne nécessite aucune intervention humaine dans le processus d'identification, (2) non supervisée et ne requiére aucune connaissance a priori sur les données à analyser, (3) séparer automatiquement les spammeurs des utilisateurs légitimes, alors que les méthodes existantes nécessitent que l'utilisateur spécific empiriquement un seuil de séparation, et (4) générale dans le sens que c'est une approche qui peut être appliqués à de différentes types de média sociaux, alors que certaines approches existantes sont exclusivement désignées à des applications spécifiques. Nous avons démontré empiriquement l'efficacité de l'approche proposée sur des données réelles extraites à partir de Instagram et Twitter.

Mots-cls—Réseaux sociaux, détection des spammeurs, distribution de Dirichlet, maximum de vraisemblance, l'algorithme EM.

[Cette page a été laissée intentionnellement blanche]

ABSTRACT

The popularity of online social networks makes them a convenient platform for malicious users such as social spammers. Hence, identifying and suspending the social spammers is decisive to preserve the social media from unsolicited content and activities. A variety of approaches have been proposed to tackle social media spammers. However, the majority of existing methods are supervised and thus mainly dependent on training data. In our investigation of current literature, we found only a few numbers of unsupervised approaches specifically designed for identifying spammers in online social services. A common limitation of the existing unsupervised methods is their dependency on human intervention in order to set an informal threshold to detect spammers.

In this thesis, we address the problem of automatic detection of spammers in online social networks. Specifically, we propose a general unsupervised approach which is capable of automatically discriminating between spammers and legitimate users in various kinds of social platforms without any necessity to prior knowledge about the data under investigation. Our approach is a principled statistical framework based on Dirichlet mixture model which is one of the most powerful probability distributions in data modeling. In this regard, the social behavior and interaction of a user with other participants is represented by a feature vector with several attributes. Next, the users' feature vectors are modeled as a mixture of Dirichlet distribution with several components. Each component in the mixture model represents a group of users with similar feature vector values which means a group of users with similar social behaviors and interactions. Then, the probability density function is estimated and the Dirichlet component that corresponds to spammers is identified. The efficacy of our proposed approach has been proved through several experiments conducted on real data extracted from Instagram and Twitter.

Keywords—Social networks, Spammers detection, Unsupervised learning, Dirichlet Mixture Model, Maximum likelihood, EM algorithm.

[Cette page a été laissée intentionnellement blanche]

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

[Cette page a été laissée intentionnellement blanche]

LIST OF TABLES

Table Page

[Cette page a été laissée intentionnellement blanche]

ABREVIATIONS

AdaBoost	Adaptive Boosting
ADTree	Alternating Decision Tree
API	Application Programming Interface
CD	Correction Detection
CDF	Cumulative Distribution Function
DT	Decision Tree
EM	Expectation Maximization
FA	False Alarm
FCM	Fuzzy C-Mean
HITS	Hyperlink-Induced Topic Search
K-NN	K-Nearest Neighbor
LogitBoost	Logistic Boosting
ML	Maximum Likelihood
MultiBoost	Multi-purpose Boosting
NB	Naive Bayes
OSN	Online Social Network
PCA	Principal Component Analysis
RBF	Radial Basis Function
SNA	Social Network Analysis
SVM	Support Vector Machine
UGC	User Generated Content
URL	Uniform Resource Locator

[Cette page a été laissée intentionnellement blanche]

CHAPTER I

INTRODUCTION

1.1 Overview

One of the primary reasons for communication evolution in our era is certainly engaged with the concept of social media. Users all over the world find and organize contacts through various kinds of social media platforms to create, share, and exchange information. Currently, social media include various forms of technologies such as Internet forums, social blogs, microblogging platforms, wikis, social networking sites, media-sharing services, and social bookmarking (Baruah, 2012).

Over and above all types of social media platforms, online social networks present an innovative way for people to interact, which differ significantly from the former networks like Web (Mislove, 2009). Facebook and LinkedIn are examples of popular online social networks used to make connections. Users join these networks to publish their own content and create links to other users in the network. Other online social networks such as Instagram and YouTube are used to share multimedia content with people, and others such as Twitter and Tumblr are microblogging sites that let users share their opinions.

In fact, the emergence of social networking sites leads to a widespread volume of user generated contents that spread quickly and extensively through the online media. Consequently, social networks have become a convenient target for opportunist users such as social spammers to take advantage and spread unwanted and illegitimate information. Spamming activities have grown considerably over the years and resulted in significant wasted network bandwidth and decreased quality of the service in social media. Spamming activities not only

pollute the content contributed by normal users and results in bad user experiences, but also can deceive or even trap legitimate users (Tan et al., 2012).

By now, online social networks have been abused by different forms of spamming activities. Spammers spread irrelevant content through social media in different ways such as spam messages in emails, spam blog entries, spam comments and spam posts. Generally, spammers attempt to send unwanted invitations and friend requests, promote products, start viral marketing, spread fads, and in some cases harass legitimate users of social networks in order to decrease their trust in the particular service (Bhat and Abulaish, 2013). Hence, identifying and suspending the spammers' activities is of a great importance to maintain high-quality services.

To help identify the potential spammers, the concept of social network analysis can be leveraged. (Bouguessa, 2011) defines social network analysis as a task which outlines the interactions between users and/or group of users and their resources with the objective of understanding their behavior and intent. Thus, the notion of social network analysis not only establishes an in-depth perspective of social network structure but also is a convenient source of information to develop the algorithms that are capable of detecting various types of users such as anomalous or influential users.

The literature shows a variety of approaches to tackle the problem of spammers in online social networks. The spammer detection approaches can be broadly divided into two main categories: supervised approaches and unsupervised approaches. Typically, in supervised approaches (Lee et al., 2010), (Benevenuto et al., 2009), (Bhat and Abulaish, 2013) users' activities on the social media platform are presented by defining a set of features. Then, the defined features are used as attributes of supervised machine learning algorithms to classify users as either spammers or legitimate users. In the case of supervised machine learning algorithms, a comprehensive and representative training data is required. There are also other supervised approaches that apply supervised matrix factorization (Zhu et al., 2012) or online learning (Hu et al., 2014) to detect spammers.

Different from supervised methods, unsupervised approaches Tan et al. (2013), (Bouguessa,

2011) identify spammers by searching for coherent structure in an entirely unlabeled data without relying on the training data. Due to the fact that collecting unlabeled data for unsupervised approaches is relatively easier than gathering labeled training data for supervised methods, there is a good reason to focus on unsupervised approaches. As a deduction, the purpose of this thesis is introducing a principled unsupervised approach to identify spammers in online social networks.

1.2 Motivations

Even though a variety of approaches has been proposed to detect spammers in online social networks, this area still offers opportunities for further improvements. Existing supervised approaches are capable of identifying spammers in many types of online social networks (Benevenuto et al., 2008), (Lee et al., 2010) but as it was mentioned before they are largely dependent on training data whereas gathering an inclusive labeled training set for supervised machine learning algorithms is expensive, is time consuming and requires an extensive human effort.

In the case of unsupervised approaches, only a few publications (Viswanath et al., 2014), (Tan et al., 2013), (Bouguessa, 2011) can be found that discuss the specific problem of detecting spammers in online social networks, which means there are still relevant problems to be addressed. For example, a key limitation of some recent unsupervised approaches (Tan et al., 2013), (Viswanath et al., 2014) is their dependency on a user-specified threshold to detect spammers. In real situations, however, it is rarely possible for users to supply the threshold values accurately. Spammer detection accuracy can thus be seriously reduced if an incorrect threshold value is used. In addition to this, in any case, the optimal threshold depends on the spammer detection algorithm being used and there is no single threshold suitable for all purposes.

The work described in (Viswanath et al., 2014) introduces an unsupervised approach based on principal component analysis (PCA). Despite the fact that, one of the PCA's intuitions is coming from the Gaussian distribution (Nie et al., 2014), the method may not be conventional in case of analyzing non-Gaussian data whereas social networks data are non-Gaussian with

non-symmetric shapes. (Bouguessa, 2011) proposed a statistical framework based on the beta mixture model to identify spammers in a university-scale email network by estimating the communication reciprocity of the users. Thus, the approach is limited to one-dimensional data since it only considers the communication reciprocity of the users. Also, it should be mentioned that the approach in (Bouguessa, 2011) is designed for a specific kind of social network in which interactions flow from an initiator to a receiver (e.g. email networks).

Other existing approaches (Narisawa et al., 2006), (Uemura et al., 2008) consider unsupervised methods for the detection of spam documents from a given document sets. It is important to note that these techniques are mainly focused on filtering spam contents and not primarily designed to detect social spammers since they do not consider users' social activities and behaviors which, in turn, limit their applicability to analyze social network data.

1.3 Contribution

The purpose of this thesis is to alleviate the aforementioned limitations of existing unsupervised approaches by developing a novel and principled approach for detecting spammers in online social systems. In a nutshell, our approach starts first by representing each user of the social network with a feature vector that reflects its behavior and interactions with other participants. Next, we propose a statistical framework based on the Dirichlet mixture in order to model the estimated users' feature vectors. The probability density function is therefore estimated and the Dirichlet component that corresponds to spammers is identified.

Note that we have used the Dirichlet distribution mainly because it permits multiple modes and asymmetry, and can thus approximate a wide variety of shapes (Bouguila et al., 2004), (Ma and Leijon, 2009), while several other distributions are not able to do so. The use of the popular Gaussian distribution, for example, may lead to inaccurate modeling (e.g. overestimation of the number of components in the mixture, increase of misclassification errors, etc.) because of its symmetric shape restriction (Boutemedjet et al., 2010). Due to the limitations of the Gaussian distribution, we believe that this distribution could not be used to cluster the users' feature vectors into several components. The number of components

in the mixture will be over-estimated and the identification of spammers will be, in turn, not obvious.

To summarize, in contrast to several distributions, the Dirichlet distribution is more flexible and powerful, since it permits multiple symmetric and asymmetric modes; it may be skewed to the right, skewed to left or symmetric (Bouguila et al., 2004). This great shape flexibility of the Dirichlet distribution provides a better fitting of the users' feature vectors, which leads, in turn, to a substantially improved modeling accuracy.

The significance of our work can be summarized as follows:

- We propose a principled approach based on the Dirichlet mixture model to automatically identify spammers in online social services. To the best of our knowledge, this work represents the first application of the Dirichlet mixture model to social network data.
- The proposed method is parameterless and does not require any prior knowledge about the data under investigation while existing unsupervised approaches require human intervention in order to set informal thresholds to detect spammers.
- The proposed method is general in the sense that it can be applied to different social online services since it exploits several user behavior-based features while, as previously mentioned, some existing unsupervised approaches such as (Tan et al., 2013) and (Bouguessa, 2011) deal only with specific types of online services.
- We conducted experiments on real data extracted from different online social sites such as Instagram and Twitter. The experimental results suggest that the performance of our unsupervised approach is comparable to (and, in some cases, even better than) those of supervised techniques that have the advantage of using labeled data.

1.4 Thesis Plan

The rest of this thesis is organized as follows. Chapter 2 provides a comprehensive definition of social media, principal concepts and general methodologies of social network analysis

and social spammer phenomena. Chapter 3 explains background information and provides a literature review on spammer detection approaches in online social networks. Chapter 4 describes the proposed approach in detail. Chapter 5 is devoted to the presentation of our experimental results and the evaluation of our proposed approach compared to several supervised machine learning-based algorithms. Finally, we present the conclusions of work in Chapter 6.

CHAPTER II

BACKGROUND

This chapter provides a comprehensive definition of social media in Section 2.1 and online social networking services in Section 2.2. The concept of social network analysis is discussed in Section 2.3. A brief history of applying data mining techniques for analyzing online social network data is explained in Section 2.4. Afterward, the principal concepts and general explanation of several machine learning methods is provided in Section 2.5. Finally, social spammer phenomenon is discussed in Section 2.6.

2.1 Social Media

Since the beginning of the Internet, there have consistently been various kinds of information sharing networks, as the most wildly known of which is the World Wide Web (Mislove, 2009). The early stages in the World Wide Web are engaged with the concept of Web 1.0 which is the first generation of web. Web 1.0 is considered read-only web for only broadcasting information to the users. Users were allowed to search the information and read it with a very limited user interactions or content contributions. In fact, users had access to web pages, but they were passive viewers that could not contribute to the content of the web pages (Aghaei et al., 2012).

By the emergence of Web 2.0 sites, users were allowed to interact and collaborate with each other as creators of user-generated content in online communities instead of just retrieving information. Users created an account on the Web 2.0 sites and they were able to collaborate. For example, they were able to comment on the published articles. Web 2.0 includes a variety of services such as social networking sites, self-publishing platforms, personal websites,

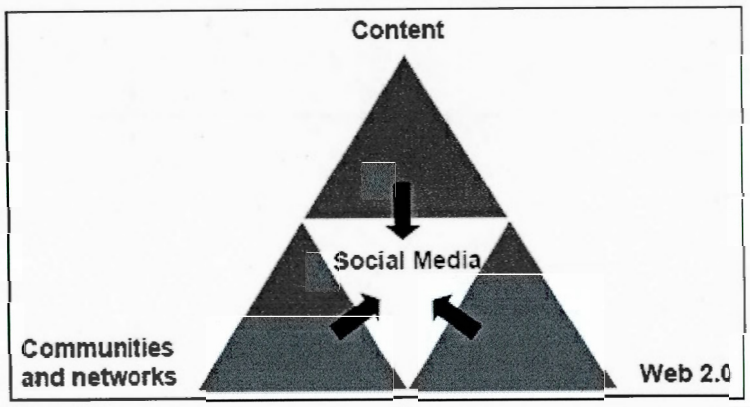

Figure 2.1: Social media triangle.

blogs, wikis, media-sharing sites, tagging services, hosted services, web applications, social bookmarking, etc. Architecture of participation is the term applied to the structure of these sites in which users are capable of participating in the content of an application as they work with it (O'Reilly, 2007).

Based on the ideological and technological foundation of Web 2.0 that allows creation and exchange of user-generated content (Shewmaker, 2014), the concept of social media emerged in referring to the means of interaction among people in which they generate, share and exchange information, ideas and contents in virtual communities and networks. The social media phenomena (e.g. Facebook and Twitter) is considered one of the most prominent developments in the Internet world in recent years (Ahlqvist et al., 2008) which also has become the most important platform for people to seek and exchange information (Wang and Lee, 2014). As depicted by Figure 2.1 [1], the core of social media is based on content, user communities and Web 2.0 technologies. Based on (Kaplan and Haenlein, 2010) classification, social media takes on six different types: (1) collaborative projects, (2) blogging and microblogging services, (3) content communities, (4) social networking services, (5) virtual games and (6) virtual social worlds.

[1] http://wordpress.viu.ca/cstewart/2014/09/15/what-is-social-media/

2.2 Online Social Networking Services

According to (Boyd and Ellison, 2007), "social network sites can be defined as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system". The characteristics and definition of these connections depend on the nature of each social network. Among all types of social networks, online social networking services are the most popular sites on the Internet. Users of these sites form a social network, which provides a powerful means of sharing, organizing, and finding content and contacts. Facebook and LinkedIn are examples of the online social networking sites used to find and organize contacts. Other social networks such as Flickr, YouTube, and Instagram, are used to share multimedia contents, and others such as BlogSpot and Twitter are used to share blogs and microblogs (Mislove, 2009).

2.3 Social Network Analysis

Due to the fact that social network sites are conveniently accessible through the Internet and Web 2.0 technologies, users are becoming more involved in social networks to acquire information, news and opinions of other users on different topics. According to (Adedoyin Olowe et al., 2013), "social networks are important sources of online interaction and content sharing, subjectivity, assessments, approaches, evaluation, influences, observations, feelings, opinions and sentimental expressions borne out in text, reviews, blogs, discussions, news, remarks, reactions, or some other documents". In this regard, network analysis (SNA) can be effectively applied in terms of studying social networks, collaboration structures and new types of social interactions for such a large-scale data.

According to (Pinheiro, 2011), the concept of social network analysis has been proposed in referring to the use of network theory to analyze social networks. Social network analysis (SNA) analyzes social relationships by means of network theory wherein nodes represent, for example, users of the social network and ties represent relationships between them (Abraham et al., 2009). In other words, social network analysis is a set of theories, tools, and processes for understanding the social network structure and users' relationships. Social network analysis

practitioners analyze the collected network data to illustrate the patterns of connections between the users of the network.

2.4 Data Mining Techniques and Social Network Analysis

Fast-paced information exchange among users of online social networks generates massive data characterized by three computational issues: size, noise and dynamism (Adedoyin Olowe et al., 2013). In order to analyzing large-scale data of the social networks within a reasonable time, an automated information processing is required. Social network sites are perfect sources of information to mine with data mining tools, since data mining techniques require huge data sets to be able to mine patterns from data. In this regard, data mining techniques seems to be a suitable tool for discovering valuable, accurate and relevant knowledge from social network data.

Data mining provides a wide range of techniques to acquire useful knowledge such as trends, patterns and rules from massive data sets (Kagdi et al., 2007). Data mining techniques are also used for information retrieval, statistical modeling and machine learning. These techniques consist of several procedures such as data pre-processing, data analysis, and data interpretation processes.

There are various types of data mining techniques for sanalyzing social network data such as graph theoretic, community detection, topic detection and tracking (TDT), etc. In the early stages of social networks, graph theory is probably the main method for analyzing social network data. In the field of graph theory, important features of the network such as the nodes and links are identified for better understanding of the characteristics of the network. Community detection is an approach which mostly employ hierarchical clustering to group the nodes of the network in order to detecting individual communities. Topic detection and tracking (TDT) is a method in which new topics (or events) of the social network are identified in order to tracking their subsequent influences over a period of time.

Dynamic analysis and static analysis are important topics in data mining. Static analysis is an easier task in comparison with the dynamic analysis for streaming networks. In static analysis, the network analysis is performed in batch mode since it is considered that social

network changes gradually over time. Conversely, dynamic analysis for streaming networks such as Facebook and YouTube is more complicated since data on these networks are generated at high speed and capacity. Dynamic analysis of social networks is the subject of several studies such as interactions between entities (Papadopoulos et al., 2012), temporal events on social networks (Becker et al., 2011), and evolving communities (Fortunato, 2010).

2.5 Machine Learning in Social Media Analysis

The purpose of machine learning is studying and constructing the algorithms that are capable of learning from data. These algorithms aim at finding patterns and making predictions from data and are involved with various concepts such as multivariate statistics, pattern recognition, and advanced predictive analytics. One of the most important application of machine learning algorithms is in the field of data mining (Kotsiantis, 2007).

While working with large, diverse and fast changing data sets, machine learning methods are very effective in finding out important predictive patterns which are to be discovered from data. Furthermore, in the case of social network data, machine learning shows improvements in terms of accuracy, scale and speed compared to traditional methods. In social network analysis, machine learning can be applied to discover and analyze the communities, social activities and interactions, user behavior, etc.

According to (Russell et al., 1996), a general classification for machine learning approaches consists of three broad categories:

- Supervised learning: a machine learning task in which sample inputs and their desired outputs are fed into the learning algorithm, and the purpose is learning a general rule that is capable of mapping inputs to outputs.

- Unsupervised learning: a machine learning task in which the learning algorithm is not provided with labeled samples, and the goal is finding structure in the input data and looking for hidden patterns in data.

- Reinforcement learning: a machine learning task in which a computer program is in

interact with a dynamic environment for a specific purpose, without a tutor supervision (Bishop, 2006).

Machine learning techniques are applied in various fields such as classification, clustering, regression, density estimation, and dimensionality reduction. Classification is a method of supervised learning in which inputs are divided into two or more classes. The learner generates a model that assigns the label to upcoming inputs. Classification consists of a variety of algorithms such as Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (K-NN), etc. One of the domains that classification techniques can be effectively applied is spam filtering (Bishop, 2006).

Clustering is a method of unsupervised learning in which the goal is to divide the inputs into groups. Unlike classification, there is no knowledge about the groups beforehand. Clustering is a common technique for statistical data analysis. Regression is also an unsupervised task with continuous outputs rather than discrete outputs. Density estimation is an unsupervised learning that attempts to learn the underlying probability distribution. In dimensionality reduction, inputs are mapping into a lower-dimensional space for simplification. An example of dimensionality reduction is topic modeling, in which a list of human language documents is given to a program and the task is to find out similar topics in the documents.

2.6 Social Spammers Phenomena

While social media services have emerged as important platforms for information distribution and communication, it has also become infamous for spammers who overwhelm other users with unwanted content. The (fake) accounts, known as social spammers (Lee et al., 2010), (Webb et al., 2008), are a special type of spammer who matches up to launch various attacks such as spreading ads for sales; spreading pornography, viruses, or fishing; making friends with victim and illegitimately grabbing their personal information (Bilge et al., 2009); or disrupt reputation (Lee et al., 2010).

CHAPTER III

REVIEW OF THE LITERATURE

The literature on spam detection strategies shows a variety of approaches designed for different types of social media platforms such as e-mail services, social networking sites, media sharing networks, social blogging, microblogging, question-answering, social bookmarking, etc. In general, the spammers' detection approaches can be broadly divided into two main categories: supervised approaches and unsupervised approaches. In this chapter, we provide an overview of related studies within these two main categories. General concepts of supervised and unsupervised approaches are discussed in Section 3.1. Previous works regarding supervised methods are reviewed in Section 3.2. The literature on the specific problem of identifying spammers in online social networks through unsupervised approaches is reviewed in Section 3.3.

3.1 Introduction to Supervised and Unsupervised Approaches

Before the emergence and popularity of online social networks, e-mail services were the most common platform for spamming activities. Eventually, similar abuses appeared on other platforms such as Web search engine spam, spam in blogs, video spam, wiki spam, Internet forum spam, social spam, etc. In recent years, both supervised and unsupervised anti-spam strategies for email services have been widely studied to the extent that statistics shows a considerable decrease in the volume of spam emails. Hence, recent studies are more focused on methods which are capable of identifying social behavior patterns of spammers to efficiently detect social spammers in online social networks.

To identify spammers in social networks, supervised methods mostly employ machine learn-

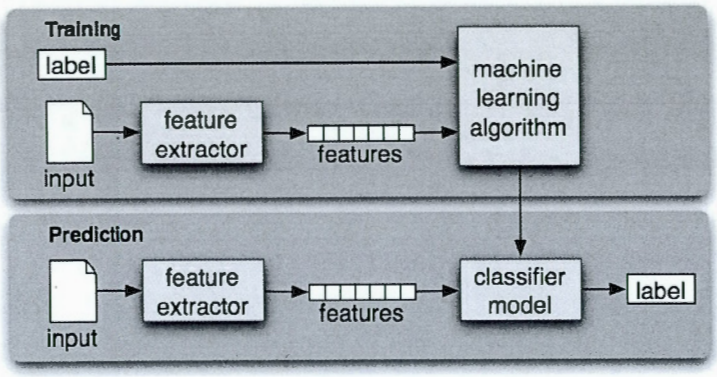

Figure 3.1: Supervised learning.

ing algorithms as classifiers. A representative training data set is thus required to learn the classifiers. The training set is generally built upon a set of features. There are various categories of features with different characteristics to discriminate between spammers and normal users. A number of studies are mainly focused on content-based features to detect spam posts while other studies employ social behavior-based features to reflect users' social behaviors on the social network. Also, a combination of all these features were studied through several approaches to achieve better results. Then, based on the characteristics of the data and selected features, the proper supervised machine learning algorithm is employed to classify users of social networks in order to identifying the class of spammers. For the purpose of illustration, the framework of supervised classification is demonstrated in Figure 3.1 [1]. During the training phase, a model is constructed based on the feature sets and their corresponding labels that are fed into the machine learning algorithm. In the prediction phase, predicted labels (e.g. spammer or legitimate user) are generated based on the feature sets are fed into the model.

Different from supervised methods, the unsupervised approaches employ algorithms that are capable of categorizing users through existing unlabeled data without relying on training data. Unsupervised algorithms are employed in order to find patterns in the input data. Classic examples of unsupervised learning are clustering and dimensionality reduction. The framework of unsupervised learning is demonstrated in 3.2. Unlabeled data is scaled to be

[1] http://www.nltk.org/book/ch06

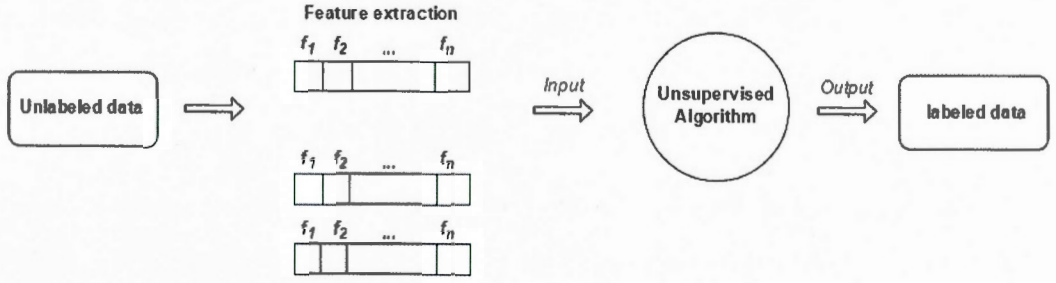

Figure 3.2: Unsupervised learning.

in the form of feature vectors and used as input to unsupervised algorithms. Similar to supervised approaches, various categories of features such as content-based, social behavior-based or combination of both features are employed to reflect the user's behavior on social networks.

3.2 Supervised Approaches to Detect Spammers

Numbers of supervised approaches have been proposed in the literature to detect spammers in social media networks. (Benevenuto et al., 2008) and (Benevenuto et al., 2009) studied the problem of detecting video spam contents on YouTube which is a video-sharing social network. The authors in (Benevenuto et al., 2008) and (Benevenuto et al., 2009) addressed the problem through a supervised approach that employs SVM machine learning algorithm as its classifiers to discriminate between promoters, spammers, and legitimate users. Both user-based and social behavior-based attributes were considered to extract a powerful set of features that detect spammers, as well as promoter.

The approach proposed by (Tseng and Chen, 2009) is a supervised method for spammer detection in email social network based on email communications. (Tseng and Chen, 2009) designed a system called MailNet that considers incremental updates to capture the evolving nature of email communication. MailNet includes two major processes, initial off-line training, and incremental on-line update. The email social network is constructed by means of a training set of emails. Then, several features are extracted from each user in the network.

Afterward, the method employs the SVM algorithm in two phases, initial SVM training and incremental update of SVM model to re-train the SVM model and achieve better results. For the purpose of evaluation, several experiments were conducted on the small university-scale email server of the Computer Center at National Taiwan University.

(Markines et al., 2009) address the problem of spammer identification in social tagging systems and more specifically in social bookmarking services. In the suggested approach, the authors focused on defining and analyzing the features which are suitable for capturing social spam. They proposed six features involved with various levels of spam activity such as post level, resource level, and user level. The set of features are mostly content-based and are related to the post and resource levels. For the experimental phase, several machine learning algorithms such as SVM, AdaBoost and RandomForest were employed. Then, to evaluate the efficiency of the features in detecting spammers, data sets from a social bookmarking service called Bibsonomy [2] were used. The approach is mainly designed to detect social spammers in social tagging systems.

A comprehensive spammer behavior analysis were performed in (Lee et al., 2010) and (Stringhini et al., 2010) inspired by the concept of honeypots (honey-profiles). According to (Stringhini et al., 2010), "the purpose of honeypots or honey-profiles is to log the traffic (e.g. friend requests, messages, invitations) they receive from other users of the network". (Stringhini et al., 2010) analyzed the collected data from honey-profiles on Facebook, Twitter, and MySpace. The authors in (Stringhini et al., 2010) identified four categories of spam bots: (1) Displayer, bots that only display spam content on their own profiles, (2) Bragger, bots that post messages to their own feed, (3) Poster, bots that send direct message to each victim, and (4) Whisperer, bots that send private messages to their victims. Then, to detect bragger and poster spammers, several features such as the message similarity, number of sent messages, number of friends, etc. were extracted. The Random Forest algorithm was employed as a classifier from the Weka [3] machine learning toolkit.

The authors in (Lee et al., 2010) employed social honeypots on MySpace and Twitter to

[2] http://www.bibsonomy.org/
[3] http://www.cs.waikato.ac.nz/ml/weka/

monitor and analyze social behaviors of spammers. Then, the collected data were analyzed to identify the anomalous behavior of spammers which tried to contact the honeypots. Based on the gathered information through honeypots, several categories of spammers were identified such as click traps, friend infiltrators on MySpace; and promoters, phishers on Twitter. Then, four broad classes of user attributes including user demographic, user contributed content, user activity features, and user connections were considered to extract discriminative features in identifying spammers. The empirical evaluation results are based on various classifiers of the Weka machine learning toolkit such as Decorate, LogitBoost, and Bagging.

(Benevenuto et al., 2010) focused on defining a robust set of features to identify spammers on Twitter. A large training set was manually gathered and labeled for typical spammers and normal users on Twitter. (Benevenuto et al., 2010) analyzed the large set of attributes which reflects user behavior of the social network, as well as characteristics of the content posted by users. In total, 39 features related to the content of tweets and 23 features based on the user behavior were considered. Then, the top ten attributes were selected based on X^2 (Chi-Squared), which is a feature selection method in (Yang and Pedersen, 1997). The non-linear SVM classifier with the Radial Basis Function (RBF) classifier were employed to identify spammers on Twitter.

(Soiraya and Thanalerdmongkol, 2012) focused on exploiting content-based features for social network spam detection. The selected features are all text-based features such as the number of keywords, the average number of words, the text length, and the number of links. Then, the data mining model using the Decision Tree J48 is created by means of the Weka toolkit. The proposed spam identification model by (Soiraya and Thanalerdmongkol, 2012) is exclusively designed for Facebook.

(Tan et al., 2012) performed a detailed research on spammers' behavior on the large scale data from a commercial search engine [4]. They provide a comprehensive statistical overview regarding non-textual behavior of spammers including posting habits, spam hosting behaviors, link patterns, etc. After studying the non-textual behavior of spammers, five set of features including user activities, post contributions, link patterns, hosting behaviors, and content

[4]Due to commercial issues, the name of the website was not mentioned.

metadata were defined. Then, the features are used as input to Naive Bayes, Logistic Regression and Decision Tree machine learning algorithms. To optimize the method for real-time detection, (Tan et al., 2012) proposed an algorithm called BARS (Blacklist-assisted Runtime Spam Detection) with the help of an auto-expanding spam blacklist, and a high priority non-spam whitelist. The non-textual behavior features are generated at runtime based on the new post and past posts of the same user. A spam URL blacklist is also maintained to help identify new spam posts. However, the algorithm is not suitable for cases where spammers use new user IDs.

(Bhat and Abulaish, 2013) studied one of the important properties of social networks which is the clustering property of users, such as the formation of user communities. According to (Bhat and Abulaish, 2013), "in a community, the nodes are relatively densely connected to each other but sparsely connected to other dense groups in the group." (Bhat and Abulaish, 2013) aim at improving spammer classification models by incorporating community-based features of users besides the basic topological features. Considering community-based characteristics, such as interaction behavior of users within and across network community structures, in the classification can make it more difficult for spammers to qualify as legitimate users. In this regard, the weighted interaction graph of the social network is used. The weight of a directed link in the graph represents the total number of messages or posts, sent from the origin to the destination. Several features such as total out-degree, total reciprocity, total in/out ratio, and community membership are exploited. Then, classification algorithms such as decision trees, Naive Bayes, and k-NN are employed in order to detect spammers in online social networks.

(Zhu et al., 2012) proposed the Supervised Matrix Factorization method with Social Regulation (SMFSR) for spammer detection in social networks which considers the social activities and relations of the users. Specifically, (Zhu et al., 2012) proposed a joint optimization model that is capable of feature extraction and classifier learning at the same time. Then, a matrix factorization model is employed to induce a set of latent features for different users. The latent feature learning process is guided by the social relationship graph and the label information. The method is tested on data from Renren [5] which is a Chinese social networking

[5]http://www.renren.com/

service popular among college students.

(Jin et al., 2011), proposed an active learning system called SocialSpamGuard designed for real-time spam detection in social media networks. To build a convenient learning model, social network features were considered as well as content-based features. In the first phase, (Jin et al., 2011) suggested an algorithm called GAD (General Activity Detection) for fast clustering on large data to classify a large scale of historical samples into diverse clusters in order to practically generate a labeled training pool. In the second phase, based on the developed classification model, upcoming online activities will be monitored in a real-time procedure. For each new instance, the system makes predictions based on the trained model, if it is uncertain the instance is sent for human labeling and will be added to the training pool otherwise the instance will be classified.

(Hu et al., 2014) proposed a framework for social spammer detection on Twitter based on online learning. In this platform content and social network information are modeled separately, and then a unified framework is proposed to integrate both information. To model social network information, a variant of directed graph Laplacian is employed. Then, instead of learning the word-level knowledge, (Hu et al., 2014) proposed to model the content information from topic-level based on a non-negative matrix factorization model (NMF) and both models were integrated together for online social spammer detection.

Despite the promising results in detecting spam content and spammers through the aforementioned supervised approaches, their high dependency on the training data is not ignorable. The key limitation is that gathering a labeled training set is an expensive and time-consuming task. Moreover, since spammers constantly modify their spamming patterns, thus the labeled data needs to be updated constantly and the classification models need to be re-learned.

3.3 Unsupervised Approaches to Detect Spammers

Former research regarding unsupervised approaches were more focused on detecting spam content. The unsupervised approach discussed in (Yoshida et al., 2004) is exclusively designed for email systems. (Yoshida et al., 2004) decided to focus on the e-mail servers due to the possible accessibility to the extensive volume of e-mail traffic. The proposed approach

introduced an unsupervised learning engine which uses document space density information with a short whitelist to detect spam messages. However, the approach is conventional only for e-mail systems and designed for detecting spam messages.

(Narisawa et al., 2006) proposed an unsupervised algorithm to identify spam content in blogs based on the vocabulary size of blog entries which is the number of strings with the same frequencies. This method is considered a content-based analysis of spam posts in blogs since for the most part it focuses on the vocabulary size of spam entries. (Narisawa et al., 2006) noticed that spam posts' vocabulary size amplified abnormally due to the existence of extensive copies of them in blogs. However, this method is not entirely capable of detecting spam entries, specifically when the spam post is infrequent.

The authors in (Narisawa et al., 2007) proposed an unsupervised approach that is focused on identifying spam content through syntactic analysis and equivalence relations of strings. The proposed method detects spam posts based on the irrelevancy of the substrings in user generated content. In fact, (Narisawa et al., 2007) tried to discriminate spam content from legitimate content by calculating the deviations in substring frequencies of documents. (Narisawa et al., 2007) found a threshold value which separates the spam part from non-spam by means of a heuristic method and model them using a linear model. Then, the point of separation where two linear models best explain the data points is identified.

(Uemura et al., 2008) proposed an unsupervised method in order to identify a specific type of spam messages called blog spams. The method addressed the problem of detecting spam documents from a mixture of spam and non-spam documents with the concept of document complexity. (Uemura et al., 2008) believed that spam documents have less document complexity compared to normal posts in the blogs. Therefore, the proposed algorithm called DCE (Document Complexity Estimation) was developed to estimate the document complexity by means of suffix trees.

(Zhu et al., 2011) proposed a framework which monitors the returned online search results to detect spam blogs from search engines. The method monitors the top-ranked results of a sequence of temporally-ordered queries and detects Splogs (spam-blogs) based on the temporal

behavior. The temporal behavior of a blog is maintained in a blog profile. Then, based on the blog profile, the Splog (spam-blogs) detection function is employed to detect spam-blogs. Even with acceptable results, the proposed method is practical in detecting popular Splogs that have successfully passed the spam filters and are actively generating spam posts.

The authors in (Bosma et al., 2012) suggest a framework for spam detection based on user spam reports. Social networking sites offer users the option to submit user spam reports for a given message, indicating a message is inappropriate. (Bosma et al., 2012) instantiated the framework in three models that introduce propagation between messages reported by a same user, messages authored by a same user, and messages with similar content. The spam detection framework is based on HITS (Hyperlink-Induced Topic Search) which is a link-structure analysis algorithm that uses the links between messages and other objects to propagate spam scores. Hence, the method is practical in cases that all users of the online platforms participate in sending reports.

Though the aforementioned approaches are considered unsupervised, however, they are mainly designed for the detection of spam documents from a given document set. It is important to note that, for the most part, these techniques are focused on filtering spam content thus they are not capable of detecting social spammers.

Considering that the previous studies do not investigate social behavior and interactions, they are not applicable for the analysis of social network data. (Tan et al., 2012) has also mentioned that purely content-based approaches encounter difficulties in detecting social spammers since malicious users exhibit unique non-textual patterns in online social networks and constantly change the content of spam messages. In our investigation of the current literature, we found a few number of unsupervised approaches that are specifically designed for detecting spammers in online social networks.

The unsupervised method suggested by (Viswanath et al., 2014) is a statistical approach which is based on Principal Component Analysis (PCA) in order to detect the irregular behaviors of users that significantly differ from the normal behaviors. To this end, the social network is presented as a matrix in such a way that rows correspond to users and columns

correspond to a set of estimated social behavior-based features. The proposed PCA-based method is employed to extract principal components through the rows of the matrix. The top-K principal components capture normal behaviors of the users whereas the rest of the components capture irregular behaviors and noise. Then, to discriminate between anomalous behaviors and noise, the bound on the L^2 norm is computed. (Viswanath et al., 2014) defined a user-entered threshold to detect anomalous users, in such a way that any user whose L^2 norm exceeds that threshold is flagged as an anomalous user.

According to Han and Liu (2014), although the PCA method is model free as a procedure, its theoretical and empirical performances rely on the distributions. With regard to the empirical concern, the PCA's geometric intuition originates from the major axes of the contours of constant probability of the Gaussian. Moreover, Han and Liu (2014) have also shown in their seminal work that this intuition did not hold in the case of non-Gaussian data. As a matter of fact, social network data are non-Gaussian, in a way that they are in from of skewed data with non-symmetric shapes. In fact, it has been also empirically shown in (Wilson et al., 2009) that, interaction activity on social network is significantly skewed towards a small portion of each user's social links. Consequently, a PCA-based approach may not be practical while the data under investigation are away from the Gaussian distribution which permits a symmetric bell shape only. We can thus surmise that the PCA-based approach proposed by Viswanath et al. (2014) is not capable of effectively detecting spammers in online social networks since PCA encounters difficulties in modeling the complex non-Gaussian data.

(Tan et al., 2013) proposed an unsupervised scheme called UNIK (UNsupervised socIal networK spam detection) to detect spammers on social networks. First, based on the posted URLs by the users, a user-link graph is constructed. Then, a social graph is built up based on the mutual relationship of the users. Next, UNIK leverages the social graph to identify non-spammers. Based on these legitimate users, the algorithm constructs a URL whitelist containing URLs posted by the identified normal users. The whitelist is then used to filter out URL edges in the user-link graph. To effectively identify spammers, the authors in (Tan et al., 2013) propose to compute the nodes degree in the trimmed graph and flag users whose degree is beyond an input threshold as spammers. It is clear that UNIK depends heavily

on the user-link graph constructed from the URLs posted by participants. We can surmise thus that it is likely that UNIK misses potential spammers who, instead of posting URLs, adopt different strategies to spam the system. Here, we believe that the application of UNIK is limited to the social platforms in which spamming activities revolve around sending spam URLs.

(Bouguessa, 2011) proposed an unsupervised method to detect spammers in social networks in which interactions flow from an initiator to a receiver. The users interactions are modeled as a directed graph in which users are considered as nodes and the direction of the messages in the social network is indicated by the arcs direction. Next, a legitimacy score is estimated for each node by means of communication reciprocity metric. The estimated scores are then modeled as a mixture of the beta distribution to identify the beta component which corresponds to spam senders. The empirical results show that the approach is able to detect spammers on the social platform where the sender initiates an interaction with the receiver (e.g. email networks). A limitation key of the suggested method is that it considers only one feature (communication reciprocity), hence the application of the model is limited to one-dimensional data. Furthermore, we note that the proposed approach is suitable for social services such as e-mail systems and may not be applicable to other platforms.

The purpose of this thesis is to alleviate the aforementioned limitations of existing unsupervised approaches by developing a novel and principled approach for detecting spammers in online social networks. In a nutshell, our approach starts first by representing each user of a social network with a feature vector that reflects its behaviors and interactions with other participants. Next, we propose a statistical framework based on the Dirichlet mixture in order to model the estimated users' feature vectors. The probability density function is therefore estimated and the Dirichlet component that corresponds to spammers is identified.

Note that we have used the Dirichlet distribution mainly because it permits multiple modes and asymmetry and can thus approximate a wide variety of shapes (Ma et al., 2014), (Bouguila et al., 2004), while several other distributions are not able to do so. The use of the popular Gaussian distribution, for example, may lead to inaccurate modeling (e.g. overestimation of the number of components in the mixture, increase of misclassification errors, etc.) because

of its symmetric shape restriction (Boutemedjet et al., 2010). Due to the limitations of Gaussian distribution, we believe that this distribution could not be used to cluster the users' feature vectors into several components. The number of components in the mixture will be over-estimated and the identification of spammers will be, in turn, not obvious.

To summarize, in contrast to several distributions, the Dirichlet distribution is more flexible and powerful since it permits multiple symmetric and asymmetric modes, it may be skewed to the right, left or symmetric (Bouguila et al., 2004). This great shape flexibility of the Dirichlet distribution provides better fitting of users' feature vectors, which leads, in turn, to a substantially improved modeling accuracy.

CHAPTER IV

THE PROPOSED SPAMMERS IDENTIFICATION APPROACH

In this chapter, we introduce our proposed statistical framework which is based on the Dirichlet mixture model to identify spammers in online social networks. First, each user of the social network is presented by a feature vector that reflects his/her social behavior and interactions with other participants. Then, the normalized feature vector is estimated to fit the Dirichlet mixture model. The normalization process is explained in Section 4.1. Next, the application of the mixture of Dirichlet in modeling the users' feature vectors is described in Section 4.2. The process of estimating the parameters of the mixture is discussed in section 4.3. Then, the method for identifying the number of components is described in Section 4.4. In the end, the procedure of estimating probability density function and identifying the Dirichlet component that corresponds to spammers is explained in section 4.5. Figure 4.1 provides a simple visual illustration of the proposed approach.

4.1 Problem Statement

Let $\mathbf{U} = \{U_1, ..., U_N\}$ represents the set of N users such that each user U_i is represented by D-dimensional vector $\vec{X}_i = (x_{i1}, ..., x_{iD})^T$. Each element x_{id}, $(i = 1, ..., N; d = 1, ..., D)$ of the vector $\vec{X}_i$ corresponds to a legitimacy score that would reflect the reputation level and social behavior of a user in a specific online social network. We assume that smallest feature values are related to spammers, while the largest values correspond to legitimate users. Note that, in our method, we consider different features that may help to discriminate between malicious and legitimate users. For example, to identify spammers in Twitter, we will consider features such as the followers to following ratio, average time between tweets, the ratio of the number of URL posted to the total number of tweets, etc. It is clear that the values of these features

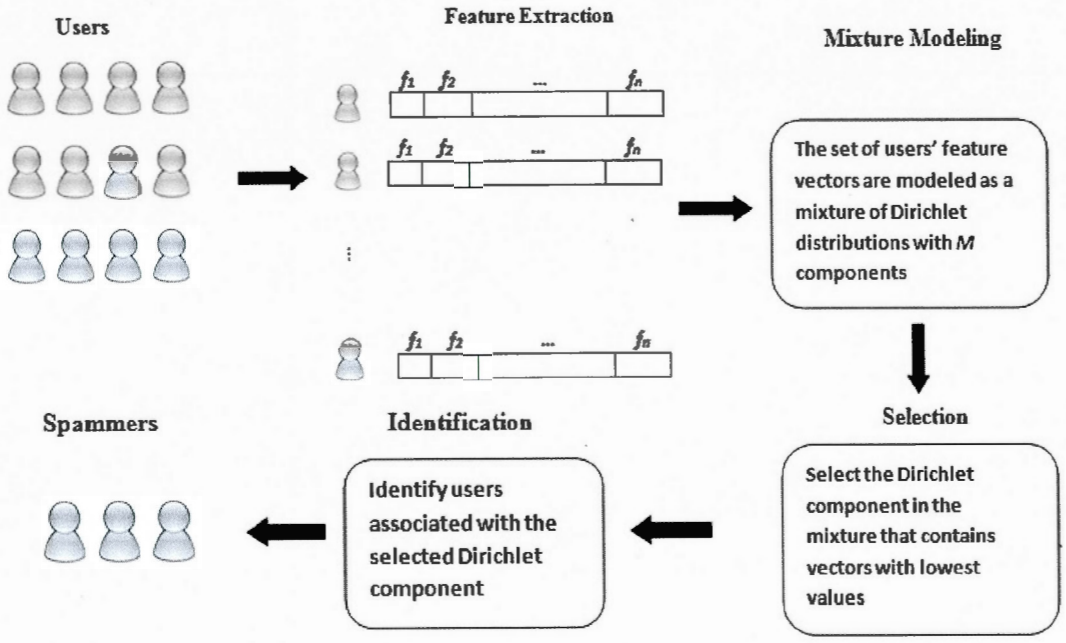

Figure 4.1: Workflow of the proposed approach.

may have different scales. In this context, it is necessary to transform these feature values into comparable, normalized values.

In our approach, we first perform log-transformation to all the estimated feature values of all users. Such log-transformation aims to squeeze together the large values that characterize legitimate users and stretch out the smallest values, which correspond to spammers. This squeezing and stretching yields comparable feature values and also contributes to enhancing the contrast between largest and smallest values. Then, to fit the Dirichlet distribution, we normalize the log transformed value of each user's feature vector in such way that the summation of all the D element of the vector $\vec{X}_i = (x_{i1}, \ldots, x_{iD})^T$ is smaller than one. Note that all along this thesis we only use the normalized values of the users' feature vectors $\{\vec{X}_i\}$.

Finally, based on the normalized users' feature vector, we propose a statistical approach which uses the Dirichlet mixture model to automatically discriminate spammers from legitimate users. Specifically, $\{\vec{X}_i\}$ can be considered as coming from several underlying probability

distributions. Each distribution is a component of the Dirichlet mixture model that represents a set of users' feature vectors which are close one to another, and all the components are combined by a mixture form. The component which contains vectors with the lowest values corresponds to spammers.

4.2 The Statistical Model

The main goal of statistical modeling is to establish a probabilistic model which can characterize the patterns of the observations, capture their underlying distributions, and describe the statistical properties of the source (Ma and Leijon, 2009). Mixture models are flexible and powerful probabilistic tools for analyzing data. The approach of mixture model assumes that the observed data is drawn from a mixture of parametric distributions. Assuming that the normalized users' feature vectors are independent and identically distributed, several parametric statistical models could be used to describe the statistical properties of $\{\vec{X}_i\}$. In this thesis we propose to use the Dirichlet mixture model.

As discussed in the introduction section, compared to previous statistical model-based methods (most of which were based on the Gaussian mixture model), the Dirichlet mixture model showed better performance (Bouguila et al., 2004). This is mainly due to shape flexibility of the Dirichlet distribution. In fact, as mentioned in (Ma et al., 2014), the Dirichlet distribution may be L-shaped, U-shaped, J-shaped, skewed to the right, skewed to the left or symmetric. Such great flexibility enables the Dirichlet distribution to provide an accurate fit of the normalized users' feature vectors.

Formally, we expect that $\{\vec{X}_i\}$ follows a mixture density of the from:

$$Dir\left(\vec{X}_i \mid \pi, \vec{\alpha}\right) = \sum_{j=1}^{M} \pi_j Dir_j\left(\vec{X}_i \mid \vec{\alpha}_j\right) \tag{4.1}$$

where Dir_j is the jth Dirichlet distribution, M denotes the number of components in the mixture, $\alpha_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jD+1})^T$ is the parameter vector of the jth component, and $\pi = \{\pi_1, ..., \pi_M\}$ represents the mixing coefficients which are positive and sum to one. The density

function of the jth component is given by:

$$Dir_j\left(\vec{X}_i \mid \vec{\alpha}_j\right) = \frac{\Gamma\left(|\vec{\alpha}_j|\right)}{\prod_{d=1}^{D+1}\Gamma\left(\alpha_{jd}\right)} \prod_{d=1}^{D+1} x_{id}^{\alpha_{jd}-1} \tag{4.2}$$

where $\sum_{d=1}^{D} x_{id} < 1$ $(0 < x_{id} < 1)$, $x_{iD+1} = 1 - \sum_{d=1}^{D} x_{id}$, $|\vec{\alpha}_j| = \sum_{d=1}^{D+1} \alpha_{jd}$ $(\alpha_{jd} > 0)$ and $\Gamma(.)$ is the gamma function given by:

$$\Gamma(\lambda) = \int_0^{\infty} y^{\lambda-1} exp(-y)\, dy; \;\; y > 0 \tag{4.3}$$

4.3 Parameters Estimation

The most central task in modeling the normalized users' feature vectors with the Dirichlet mixture model is parameter estimation. To this end, the maximum likelihood estimation approach can be used to find the parameters of the mixture model. Let $\Theta = \{\pi_1, ..., \pi_M, \vec{\alpha}_1, ..., \vec{\alpha}_M\}$ denote the set of unknown parameters of the mixture and $X = \{\vec{X}_1, \vec{X}_2, ..., \vec{X}_N\}$ the set of the normalized users' feature vectors. The likelihood function of the mixture model with M components is defined as:

$$L(\mathbf{X} \mid \Theta) = \prod_{i=1}^{N} \sum_{j=1}^{M} \pi_j Dir_j\left(\vec{X}_i \mid \vec{\alpha}_j\right) \tag{4.4}$$

When computing the maximum likelihood estimates of the parameters in the mixture model, the Expectation Maximization (EM) algorithm is generally applied (Figueiredo and Jain, 2002). Accordingly, we augment the data by introducing M-dimensional indication vector $\vec{Z}_i = (z_{i1}, ..., z_{iM})^T$ for each vector $\vec{X}_i$. The indication vector $\vec{Z}_i$ has only one element equals 1 and the remaining elements equal 0. If the jth element of $\vec{Z}_i$ equals 1, that is, z_{ij}=1, we assume that $\vec{X}_i$ was generated from the jth component of the mixture.

$$z_{ij} = \begin{cases} 1 & \text{if } \vec{X}_i \text{ belongs to componet } m \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

Let $\mathbf{Z} = \{\vec{Z}_1, ..., \vec{Z}_N\}$ denote the set of indication vectors. The likelihood function of the complete data is given by:

$$L_c(X, Z \mid \Theta) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[\pi_j Dir_j \left(\vec{X}_i \mid \vec{\alpha}_j\right)\right]^{z_{ij}} \tag{4.6}$$

Usually, it is more convenient to work with the logarithm of the likelihood function which is equivalent to maximizing the original likelihood function. The log-likelihood function is given by:

$$logL_c(X, Z \mid \Theta) = \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \left[log\left(\pi_j\right) + log\left(Dir_j\left(\vec{X}_i \mid \vec{\alpha}_j\right)\right)\right] \tag{4.7}$$

From this perspective, the EM algorithm can be used to estimate Θ. Specifically, the algorithm iterates between Expectation and Maximization step in order to produce a sequence of estimates $\{\hat{\Theta}\}^{(t)}$, $(t = 0, 1, 2, ...)$, where t denotes the current iteration step, until the change in the value of the log-likelihood in (4.7) becomes negligible. Detail of each step is given below.

In the Expectation step, each latent variable z_{ij} is replaced by its expectation:

$$\hat{z}_{ij}^{(t)} = \frac{\hat{\pi}_j^{(t)} Dir_j \left(\vec{X}_i \mid \vec{\tilde{\alpha}}_j\right)}{\sum_{k=1}^{M} \hat{\pi}_k^{(t)} Dir_k \left(\vec{X}_i \mid \vec{\tilde{\alpha}}_k\right)} \tag{4.8}$$

In the Maximization step, the set of parameters $\Theta = \{\pi_1, ..., \pi_M, \vec{\alpha}_1, ..., \vec{\alpha}_M\}$ that maximize the log-likelihood are calculated given the values of $\hat{z}_{ij}$ estimated in the Expectation step.

Specifically, the mixing coefficients are calculated as:

$$\hat{\pi}_j^{(t+1)} = \frac{\sum_{i=1}^{N} \hat{z}_{ij}^{(t)}}{N}, \quad j = 1, ..., M \tag{4.9}$$

Let us now focus on estimating the parameters $\vec{\alpha}_j = (\alpha_{j1}, ..., \alpha_{jD+1})^T; \{j = 1, ..., M\}$. The values of $\vec{\alpha}_j$ that maximize the likelihood can be obtained by taking the derivative of log-likelihood of the complete data with respect to α_{jd} and setting the gradient equal to zero. Thus, we obtain:

$$\frac{\partial log L_c(X, Z \mid \Theta)}{\partial \alpha_{jd}} = \sum_{i=1}^{N} \hat{z}_{ij} \frac{\partial}{\partial \alpha_{jd}} log\left(Dir_j\left(\vec{X}_i \mid \vec{\alpha}_j\right)\right) = 0 \tag{4.10}$$

By replacing $Dir_j\left(\vec{X}_i \mid \vec{\alpha}_j\right)$ by its expression given by (4.2) in (4.10) and then computing the derivative with respect to α_{jd}, we obtain:

$$\sum_{i=1}^{N} \hat{z}_{ij} \left[\psi(|\vec{\alpha}_j|) - \psi(\alpha_{jd}) + log(x_{id})\right] = 0 \tag{4.11}$$

where $\Psi(.)$ is the digamma function given by $\psi(\lambda) = \frac{\acute{\Gamma}(\lambda)}{\Gamma(\lambda)}$.

Since the gamma function is defined through an iteration, a closed-form solution to (4.11) does not exist. Therefore, the values of the parameter vectors $\{\vec{\alpha}_j\}$ can be estimated using the Newton-Raphson method. Specifically, we estimate the $\vec{\alpha}_j$ iteratively:

$$\left[\vec{\hat{\alpha}}_j\right]^{(t+1)} = \left[\vec{\hat{\alpha}}_j\right]^{(t)} - H_j^{-1} \times G_j; \quad j = 1, 2, ..., M \tag{4.12}$$

where G_j is the first derivative vector of the complete log-likelihood as follows:

$$G_j = \left(\frac{\partial log L_c(X, Z \mid \Theta)}{\partial \alpha_{j1}}, ..., \frac{\partial log L_c(X, Z \mid \Theta)}{\partial \alpha_{jD+1}} \right)^T \tag{4.13}$$

and H is the hessian matrix where the diagonal elements correspond to the second derivative of the complete log-likelihood function and the non-diagonal elements correspond to the mixing derivatives (Bouguila et al., 2004). The second derivative is given by:

$$\frac{\partial^2 log L_c(X, Z \mid \Theta)}{\partial^2 \alpha_{jd}} = \sum_{i=1}^{N} z_{ij} \left(\acute{\Psi}(|\vec{\alpha}_j|) - \acute{\Psi}(\alpha_{jd}) \right); \ d = 1, ..., D+1 \tag{4.14}$$

and the mixed derivative is given by:

$$\frac{\partial^2 log L_c(X, Z \mid \Theta)}{\partial^2 \alpha_{jd_1} \alpha_{jd_2}} = \sum_{i=1}^{N} z_{ij} \left(\acute{\Psi}(|\vec{\alpha}_j|) \right); \ d_1 \neq d_2, d = 1, ..., D+1 \tag{4.15}$$

where $\acute{\Psi}(.)$ is the trigamma function given by $\acute{\psi}(\lambda) = \frac{\acute{\Gamma}(\lambda)}{\Gamma(\lambda)} - \left[\frac{\acute{\Gamma}(\lambda)}{\Gamma(\lambda)} \right]^2$. Thus, the Hessian matrix can be defined as follows:

$$H_j = \sum_{i=1}^{N} z_{ij} \begin{pmatrix} \acute{\Psi}(|\vec{\alpha}_j|) - \acute{\Psi}(\alpha_{j1}) & ... & \acute{\Psi}(|\vec{\alpha}_j|) \\ \vdots & \ddots & \vdots \\ \acute{\Psi}(|\vec{\alpha}_j|) & ... & \acute{\Psi}(|\vec{\alpha}_j|) - \acute{\Psi}(\alpha_{jD+1}) \end{pmatrix} \tag{4.16}$$

The inverse of the Hessian matrix is calculated according to (Bdiri and Bouguila, 2012) by employing the replace form of the Hessian matrix as follows:

$$H_j = Q_j + \delta_j A_j A_j^T \tag{4.17}$$

where Q_j is a diagonal matrix given by:

$$Q_j = diag\left[-\sum_{i=1}^{N} z_{ij}\dot{\Psi}\left(\alpha_{j1}\right), ..., -\sum_{i=1}^{N} z_{ij}\dot{\Psi}\left(\alpha_{jD+1}\right)\right] \tag{4.18}$$

δ_j and A_j^T are given as follows:

$$\delta_j = \sum_{i=1}^{N} z_{ij}\dot{\Psi}\left(|\vec{\alpha}_j|\right) \tag{4.19}$$

$$A_j^T = (a_1, ..., a_{D+1})\,;\;\; a_d = 1, d = 1, ..., D+1 \tag{4.20}$$

The inverse of the Hessian matrix can now be calculated according to the theorem of matrix inverse in (Graybill, 1983) as follows:

$$H_j^{-1} = Q_j^{-1} + \delta_j^* A_j^{*T} A_j^* \tag{4.21}$$

where Q_j^{-1} is the inverse of the diagonal matrix Q_j and could be easily calculated. δ_j^* and A_j^* are estimated as follows:

$$\delta_j^* = \sum_{i=1}^{N} z_{ij}\dot{\Psi}\left(|\vec{\alpha}_j|\right)\left[\left(\dot{\Psi}\left(|\vec{\alpha}_i|\right)\sum_{d=1}^{D+1}\frac{1}{\dot{\Psi}\left(\alpha_{jd}\right)}\right) - 1\right] \tag{4.22}$$

$$A_j^* = \frac{-1}{\sum_{i=1}^{N} z_{ij}}\left(\frac{1}{\dot{\Psi}\left(\alpha_{j1}\right)}, ..., \frac{1}{\dot{\Psi}\left(\alpha_{jD+1}\right)}\right) \tag{4.23}$$

Once H_j^{-1} and G_j are estimated, we can now implement the iterative formula of the Newton-Raphson algorithm as expressed by (4.12). Note that this algorithm requires starting values

for $\{\vec{\alpha}_j\}^{(0)}$. In our implementation, we have used the method of moments estimators of the Dirichlet distribution (Bouguila et al., 2004) to define these initial values as follows:

$$\hat{\alpha}_{jd}^{(0)} = \frac{(\acute{p}_{11} - \acute{p}_{21})\,\acute{p}_{1d}}{\acute{p}_{21} - (\acute{p}_{11})^2};\ \ d = 1, ..., D, j = 1, ..., M \tag{4.24}$$

$$\hat{\alpha}_{jD+1}^{(0)} = \frac{(\acute{p}_{11} - \acute{p}_{21})\left(1 - \sum_{d=1}^{D} \acute{p}_{1d}\right)}{\acute{p}_{21} - (\acute{p}_{11})^2} \tag{4.25}$$

where $\acute{p}_{1d}$ and $\acute{p}_{21}$ are given as follow:

$$\acute{p}_{1d} = \frac{1}{N}\sum_{i=1}^{N} x_{id};\ \ d = 1, ..., D+1 \tag{4.26}$$

$$\acute{p}_{21} = \frac{1}{N}\sum_{i=1}^{N} x_{i1}^2 \tag{4.27}$$

The Newton-Raphson algorithm converges, as our estimation of α_{jd} changes by less than a small positive value ϵ with each successful iteration, to $\widehat{\alpha}_{jd}$.

The EM algorithm can now be used to estimate the maximum likelihood of the distribution parameters. Note that EM is highly dependent on initialization (Figueiredo and Jain, 2002). To alleviate this problem, a common solution is to perform initialization by mean of clustering algorithms. For this purpose we first implement the Fuzzy C-Means (FCM) algorithm to partition the set $X = \{\vec{X}_1, \vec{X}_2, ..., \vec{X}_N\}$ into M components. Then, based on such partition, we estimate the parameters of each component using the method of moment estimator for the Dirichlet distribution and set them as initial parameters of the EM algorithm. The steps of EM algorithm for Dirichlet mixture model is summarized in Algorithm 1.

Algorithm 1: EM algorithm for Dirichlet mixture model

Input : $\{\vec{X}_i\}_{(i=1,...,N)}$, M

Output: $\{\hat{\vec{\Theta}}\} = \left[\hat{\vec{\pi}}_1, ..., \hat{\vec{\pi}}_M, \hat{\vec{\alpha}}_1, ..., \hat{\vec{\alpha}}_M\right]$

begin

Initialization

Apply the FCM algorithm to cluster the data set $\{\vec{X}_i\}$ into M components;

Estimate the initial set of parameters of each component using (4.24) and (4.25);

repeat

Expectation

Estimate $\{\hat{z}_{ij}\}$ $(i = 1, ..., N; j = 1, ..., M)$ using (4.8);

Maximization

Estimate $\left[\hat{\pi}_j\right]$ $(j = 1, ..., M)$ using (4.9);

Estimate $\left[\hat{\alpha}_{jd}\right]$ $(j = 1, ..., M; d = 1, ..., D)$ using (4.12);

until *the change in* (4.7) *is negligible*;

Return $\hat{\Theta}$;

end

4.4 Estimating the Number of Components in the Mixture

The use of the Dirichlet mixture model allows us to give a flexible model to describe the users' feature vectors. To form such a model, we need to estimate M, the number of components and the parameters for each component. First, the number of components M is an unknown parameter that must be estimated. Several model selection approaches have been proposed to estimate M (Smyth, 2000), (Bouguessa et al., 2006). In this thesis, we implemented a deterministic approach that uses the EM algorithm in order to obtain a set of candidate models for the ranges of M from 1 to M_{max} (the maximal number of components in the mixture), which is assumed to contain the optimal M (Figueiredo and Jain, 2002). We employ the Integrated Classification Likelihood Bayesian Information Criterion ($ICL-BIC$) which is one of the powerful methods to identify the correct number of clusters in the context of multivariate mixtures such as Dirichlet, even when the component densities are misspecified (Peel and McLachlan, 2000). $ICL - BIC$ is given by:

$$ICL - BIC\,(m) = -2\ log(L_m) + d\ log(N) - 2\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{z}_{ij}log\,(\hat{z}_{ij}) \qquad (4.28)$$

Algorithm 2: Estimating the number of components in the mixture

Input : $\{\vec{X}_i\}_{(i=1,\ldots,N)}$, M_max
Output: The optimal number of components $\hat{M}$
begin
  for $M = 1$ **to** M_max **do**
    if $M==1$ **then**
      Estimate $\{\vec{\widehat{\alpha}}\}$ based on (4.12);
      Compute the value of ICL-BIC(M) using (4.28);
    else
      Estimate the parameters of the mixture using Algorithm 1;
      Compute the value of ICL-BIC(M) using (4.28);
    end
  end
  Select $\hat{M}$, such that $\hat{M} = arg\ min_M \Big\{\text{ICL-BIC}(M),\ M = 1, \ldots, M\_max\Big\}$;
end

Algorithm 3: Automatic identification of spammers

Input : A set $X = \{U_1, \ldots, U_N\}$ of N users
Output: A set $S = \{S_1, \ldots, S_K\}$ of K spammers
begin
  For a given online social network, estimate a feature vector $\vec{X}_i$ for each user;
  Normalize $\{\vec{X}_i\}$, as discussed in section 4.1;
  Apply Algorithm 2 to group the users into M Dirichlet components;
  Use the results of the EM algorithm to decide about the membership of $\vec{X}_i$ in each component;
  Select the Dirichlet component that corresponds to the smallest feature values;
  Identify spammers in U associated with the set of $\vec{X}_i$ that belong to the selected component and store them in S;
  Return S;
end

where L_m is the logarithm of the likelihood at the maximum likelihood solution for the investigated mixture model, and d is the number of parameters estimated. The number of components that minimize $ICL - BIC(M)$ is considered as the optimal value of M. The procedure of estimating the number of components in the mixture is summarized in Algorithm 2.

4.5 Automatic Identification of Spammers

Once the optimal number of components have been identified, we can use the result of the EM algorithm in order to derive a classification decision about the membership of $\vec{X}_i$ to each component in the mixture. In fact, the EM algorithm yields the final estimated posterior probability $\hat{z}_{ij}$, the value of which represents the posterior probability that $\vec{X}_i$ belongs to component j. We assign $\vec{X}_i$ to the component that corresponds to the maximum value of $\hat{z}_{ij}$. We thus divide the set of users' feature vectors into several components. As discussed earlier, in our approach we assume that spammers are characterized by small feature values. To identify such a component, for each component in the mixture, we compute the average of the projected feature values along each dimension. Then, we select the component with the smallest average values as our target component. Accordingly, users associated with the set of $\vec{X}_i$ that belong to such a component correspond to spammers. The steps described in Algorithm 3 have been implemented to automatically identify spammers.

CHAPTER V

EXPERIMENTAL RESULTS

In this chapter, we perform a set of experiments to evaluate the efficacy of our proposed approach in identifying social spammers. In this regard, we decided to analyze real data from two of the most popular online social networks that are designed for two different class of social interactions. First, we consider Instagram which is an online media-sharing and social networking platform. Next, we perform experiments on data collected from Twitter which is the most famous microblogging social network. This chapter is organized as follows, in Section 5.1, the experiments specifications are described. In Section 5.2, we illustrate the performance results of our experiments on Instagram data. In Section 5.3, the performance results of the experiments conducted on Twitter data are presented.

5.1 Experiment Specifications

For the purpose of evaluation, we extracted real data from Instagram and Twitter to construct representative data sets that reflects users' social behavior patterns and reputation level on each platform. Each user of each platform is represented by one feature vector composed of several attributes. These vectors are then used as an input of our proposed approach to identify social spammers.

In the next step, we built a labeled collection of users by manually classifying each user as either spammers or legitimate. These labeled samples are used as ground truth to evaluate the effectiveness of our approach. The standard metrics that we have applied for this purpose are: (1) Accuracy, which is the proportion of correctly classified users, (2) Correct Detection (CD) rate, which measures the proportion of spammers that are correctly classified as spammers,

False Alarm (FA) rate, which measures the proportion of legitimate users that are incorrectly classified as spammers, and F-measure of the spammers class, corresponding to the harmonic mean between precision and recall of the spammers class.

We planned two sets of experiments for the real data collected from Instagram and Twitter. The goal of the first set of experiments is to evaluate the detection accuracy of our proposed approach using different subsets of features. In the second sets of experiments, to demonstrate the capability of our unsupervised approach, we compared it with a variety of supervised algorithms available in the Weka machine learning toolkit. We considered (1) meta classifiers: AdaBoost, Bagging, Decorate, LogitBoost, MultiBoost, (2) tree-based classifiers: ADTree, Random Forest and (3) function-based classifiers: RBF Networks and SVM. Note that the classification experiments were performed using 10-fold cross-validation to improve the reliability of classification evaluations.

5.2 Identifying Spammers on Instagram

Currently, by means of social networks, users can create, share, link and reuse media content in large scale (Zhao et al., 2011). Hence, media-sharing social networks such as Instagram, YouTube and Flicker are of great importance in terms of services they offer and protecting user generated contents. In this thesis, we specifically aimed at investigating Instagram which is currently considered as one of the most popular social media platforms. Instagram is an online social networking service which enables its user to post both photo and video contents on the platform and also share them on the other social networks such as Facebook, Twitter and Flicker. Each user can interact with other Instagram users by following them and/or being followed by them. Social interactions of users take on different forms such as publish media in a way that followers can see them on their feed, like media posted by other users, write comments on media posted by other users, tag other users in comments, post special hashtags, send direct messages to others, etc.

All these features let the opportunist users such as spammers to abuse the service in any possible ways. For example spammers may harass other users by sending unwanted following requests, post irrelevant comments or links on the media to promote their pages, and so many

forms of decreasing the quality of service. Hence, identifying and eliminating the spammers' accounts is important to maintain a high quality service.

5.2.1 Crawling Instagram

The Instagram API provides open access to all of the public information and contents published by users. To collect data from Instagram, we used various sets of API methods to get users' public information and media posted available at Instagram API for developers [1]. Each category of API methods is designed to retrieve specific endpoint information. According to Instagram API, user endpoints are used to get basic information about a user and the most recent media published by a user. Relationship endpoints are used to acquire social network connections such as: the list of users a user follows, the list of users a user is followed by and information about a relationship to another user. Media endpoints are used to get information about a media object, search for media in a given area and get a list of what media is most popular at the moment. Comment endpoints get a full list of comments on a media object. Like endpoints get a list of users who have liked a media. Tag endpoints get information about a tag object. Location endpoints get information about a location, a list of recent media objects from a given location and search for a location by geographic coordinate.

It is important to note that information from the get/user endpoints request is always available, regardless of whether a user's account is public or private. However, information from other endpoints are not available if the user profile is private, thus based on the Instagram API policies we could only retrieve the public profiles data. Spammers also have easier access to public profiles, hence most of the spamming activities happen to public accounts. In Table 5.1 we demonstrate each endpoint requests and the corresponding retrieved information.

In order to collect data from Instagram we built a crawler for Instagram API which implements Algorithm 4 to gather required information. The crawler ran for 5 days form 22 May 2014 to 27 May 2014, collected total number of 641 users and total number of 2051 published media. Then, the collected data was manually analyzed in order to create labeled collection of spammers and legitimate users. The labeling was done by analyzing each user's

[1] https://instagram.com/developer/endpoints/

API Method	Retrieved Information
Get/User Endpoints	full name, user name, user id, biography, website, profile picture, number of media, number of followers, number of followings
Get/Relationship Endpoints	list of followers, list of followings
Get/Media Endpoints	media id, created time, number of comments, number of likes, location id, caption, tags, longitude, latitude
Get/Comment Endpoints	list of comments on a media
Get/Like Endpoints	list of users who have liked this media

Table 5.1: List of all features and API methods used.

profile and their last 20 user's posted media during the mentioned time. In result, among 641 collected profiles, 411 of them were labeled as legitimate users and 230 of them were labeled as spammers. It is important to note that, since the classification labeling process relies on human judgment, which implies examining hundreds of user profiles, we had to set a limit on the number of users in our labeled collection. Finally, it is worth nothing that, while visiting Instagram after collecting our data, we found that most of the accounts of the users that were manually labeled as spammers were deleted by Instagram administrators. This finding testifies the quality of our labeled data set.

Algorithm 4: Instagram Crawler

Input : A list of N user-name of active users on Instagram
Output: A vector $\vec{X}_i$ corresponding to each user U_i
begin
  for *each user* U_i; $i = 1$ **to** N **do**
    collect the user information
    for *each user* U_i; $i = 1$ **to** N **do**
      collect the media posted by each user
    end
  end
end

5.2.2 Analyzing Social Behavior of Instagram Users

To the best of our knowledge, there is no research in the literature regarding the spam phenomenon on Instagram. We thus performed an in-depth research on social activity patterns of users on Instagram to understand the platform and its policy in terms of user legitimacy. We studied all aspects of users' social activities on Instagram to define a set of features that may help us to discriminate between spammers and legitimate users.

We expected that each class of users contributes in an entirely different set of interactions to pursue their goals on the social media. We analyzed the collected users' profiles and the last 20 published media to extract a combination of characteristics that properly reflects the social behaviors of each class of users. In fact, our goal was to define a set of features with a discriminatory power to identify spammers as well as legitimate users on Instagram. By exploring the large set of attributes that we had gained through crawling Instagram API, we decided to consider a set of eight features that are mostly related to the social behavior of both spammers and legitimate users on Instagram. The complete set of features is represented in table 5.2.

Features
Full name length
Total number of media posted
Followers to following ratio
Proportion of bidirectional friends
Account lifetime
Average number of posts per week
Average number of likes received per post
Idle time in days

Table 5.2: List of features.

In the following we give a brief description about the features considered for the Instagram data set in this thesis.

- **Full name lenght**: We analyzed the complete set of features retrieved through Instagram API by calling get/user endpoints such as user id, user name, full name, biography, website, and profile picture. We observed that most of the legitimate users provided

their full name in their profile while in case of spammers the field of full name was blank or filled with a very short full name with high possibility of repetitive characters. Hence we decided to calculate the full name length of each user by employing a simple method that counts the number of characters for a given full name string.

- **Total number of media posted**: In our investigation of the collected profiles, we found that in contrast to legitimate users, spammers tended to post very few numbers of media on Instagram. In fact, spammers on Instagram were more interested in spamming other users rather than contributing in normal users' activities like publishing media. Total number of media posted was obtained directly by calling get/user endpoints for each user.

- **Followers to following ratio**: By analyzing the number of followers and the number of followings in the collected profiles, we observed that spammers tended to follow a large number of users while they were only followed by very few numbers of participants. We found that the followers of spammers were also spammers in most of the cases. Contrarily, legitimate users acted quite normal in following other participants. The number of follower and the number of following for the legitimate users was not surprisingly disparate. Therefore, we considered the ratio of the followers to following as an explicit feature. The number of followers and the number of following was obtained directly by calling get/user endpoints. The ratio is expected to be close to zero for spammers.

- **Proportion of bidirectional friends**: We analyzed the complete list of followers and followings of each user precisely. We figured out that in case of spammers, only a few number of users that had been followed by spammers followed them back. Contrarily, legitimate users following behaviors are reciprocal in most of the cases. We got the list of followers and followings through get/relationship endpoints and estimated the probability of a user being followed by the same user that she/he was following. The proportion of bidirectional friends is calculated as follows:

$$Bidirectional friends = \frac{|followings \bigcap followers|}{|followings|} \tag{5.1}$$

- **Account lifetime**: Instagram API provides a large list of media-related features through get/media endpoints such as: media id, created time of media, number of comments, number of likes, location id, caption, tags, longitude, and latitude. We analyzed all these information to extract the features that may help us in detecting spammers on Intsgram. While we were exploring the media posted by users we found that spammers had a very short account lifetime. In this thesis, by account lifetime we mean the period from the moment, a user published his/her first media on Instagram until the day that we were gathering the Instagram data. We noticed that in comparison with the legitimate users, spammers' profiles were created latterly based on their first media posted and they had remarkably shorter account lifetime. To calculate the account lifetime in days, we retrieved the created time of the user's first published media by calling get/media endpoints, and then the days before the day that we collected the data was counted.

- **Average number of posts per week**: We observed that in most of the cases spammers tended to randomly publish media. In fact, we found that spammers published a media on the first day that they joined Instagram and casually posted a few other media to act like normal users, while most of the legitimate users were constantly active on the platform and published media on a regular basis. In general, legitimate users tended to post media every week. In result, we consider the average number of posts per week as one of the influential features to identify spammers on Instagram.

- **Average number of likes received per post**: We focused on the social interactions between users on Instagram. We noticed that legitimate users' reputation level and contribution in social interactions were considerably different form spammers. For example, when a legitimate user publishes a media, the number of likes received by his/her followers is correspondent to the number of his/her followers. Furthermore, the number of likes received for a media can be considered as a meaningful parameter to

demonstrate the users' reputation level in a way that it reflects the satisfaction of Instagram members with the published media. Therefore, the average number of likes received per post is a discriminative feature to identify spammers on Instagram since the posted media by spammers had received very few likes compared to normal users. By calling get/media endpoints we retrieved the number of likes for each media, then by calculating total number of likes received for all of the posted media and dividing it by the total number of media, we obtained this feature.

- **Idle time in days**: As an important social behavior, we observed that spammers tended to be idle for long periods of time since the last posted media. By idle time we mean the period of time in days since the last time that a user published a media on Instagram. This parameter shows the consistency of users' contribution in social activities. In general, we found that legitimate users were not idle for more than a week whereas in most of the cases spammers were idle for long time after their few published posts. By calling get/media endpoints we could get the created time of the last published media. Then, idle time is calculated easily from the time that we had collected the data from Instagram API. It is important to note that in this thesis we assume spammers are characterized by small feature values, thus we performed a linear inversion of the estimated idle time values so that the small inverted values correspond to spammers.

  The cumulative distribution function (CDF) of full name length, total number of media posted, followers to following ratio, and proportion of bidirectional friends for both classes of spammers and legitimate users are illustrated in Figure 5.1. Likewise, the cumulative distribution function for both classes of spammers and legitimate users for account lifetime, average number of posts per week, average number of likes received per post, and idle time are presented in Figure 5.2.

Based on our collected data set there was a noticeable difference between the values of the selected features for each class of users. In all of these cases (except idle time) higher values corresponded to legitimate users whereas the lower values was a sign of a user being spammer. In other words, legitimate users tended to have more contribution in social interactions and

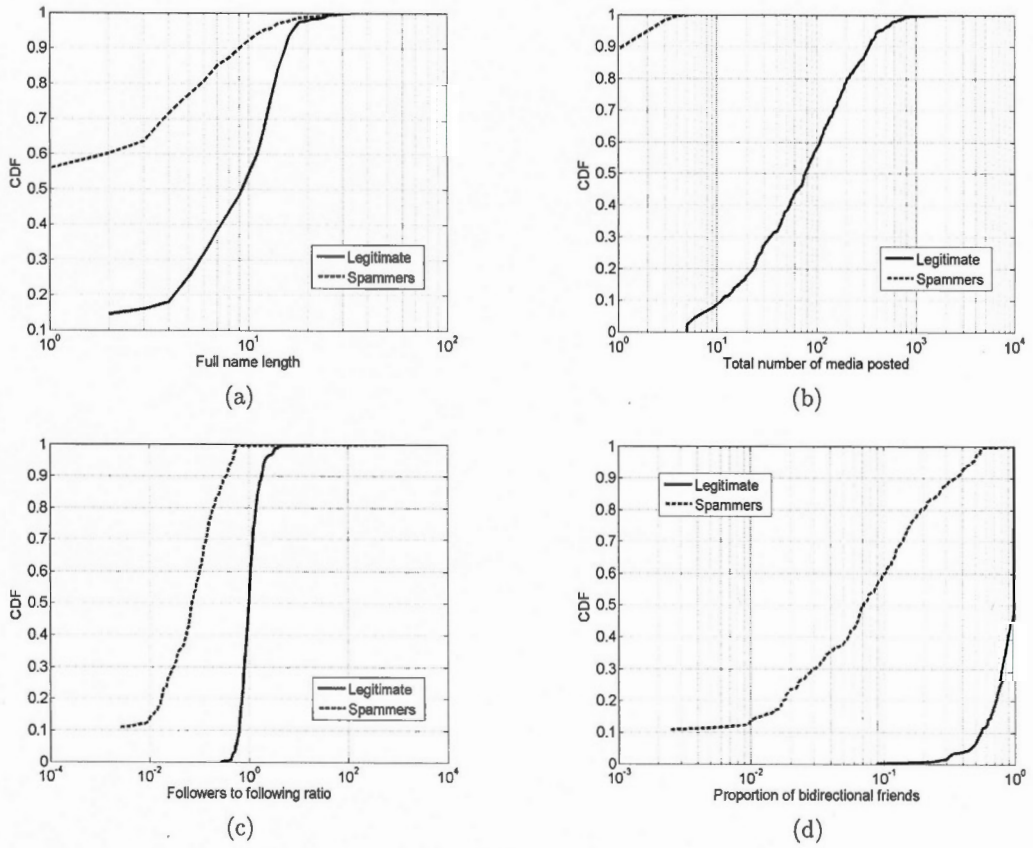

Figure 5.1: Cumulative Distribution Function (CDF) of the first four features.

built up higher level of reputation compared to spammers. As can be seen from Figures 5.1 and 5.2, in contrast to spammers, the curve for legitimate users is much more skewed toward large numbers which means they posted more media, received a lot of likes, and have more followers and bidirectional friends, while spammers exhibited quite the opposite.

5.2.3 Experiment 1

The goal of this first set of experiments was to evaluate the detection accuracy of our approach using different subsets of the eight features (full name length, total number of media posted, followers to following ratio, proportion of bidirectional friends, account lifetime, average number of posts per week, average number of likes received per post, idle time) considered in this thesis. To this end, we used the collected features of Instagram and constructed several

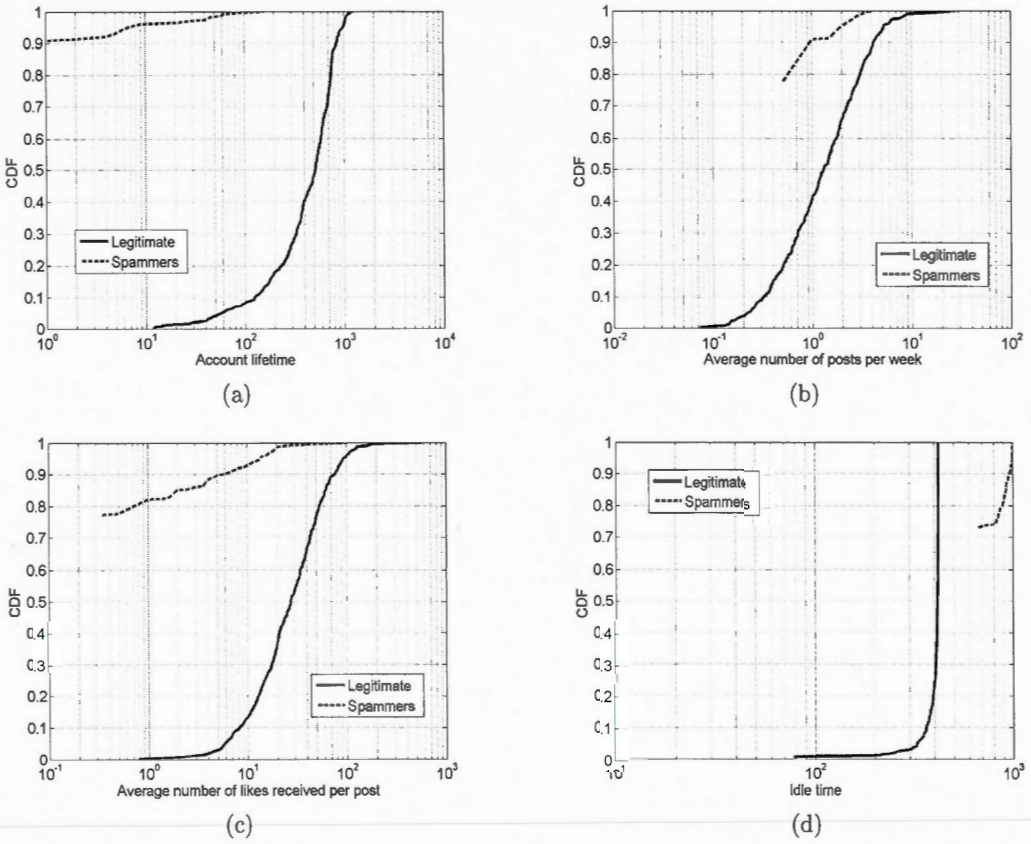

Figure 5.2: Cumulative Distribution Function (CDF) of the last four features.

data sets using the following subsets of features:

1. Proportion of bidirectional friends and full name length.
2. Proportion of bidirectional friends and total number of media posted.
3. Followers to following ratio and idle time in days.
4. Proportion of bidirectional friends and idle time in days.
5. Proportion of bidirectional friends and full name length and idle time in days.
6. Average number of likes received per post and average number of posts per week and proportion of bidirectional friends and full name length idle time in days.

7. Average number of likes received per post and average number of posts per week and proportion of bidirectional friends and total number of media posted and followers to following ratio and full name length and idle time in days.

8. The complete set of eight features.

Then, for each constructed data set, we used our approach to identify spammers. To this end, we set M_{max} to 4 (the reader should be aware that the choice of M_{max} is not limited to 4 and the user can choose other values) in all our experiments and selected the optimal number of components that minimize $ICL - BIC$. We found that the number of components varies from two to three. The Dirichlet component that represents the lowest feature values corresponds to spammers.

The created data sets differed only in the underlying features used but had the same class labels that designated spammers and legitimate users. Obviously, we have ignored the class labels when applying our approach but we used them to evaluate the detection accuracy of the proposed method. Figure 5.3 illustrates the results, evaluated with accuracy, CD rate, FA rate and F-measure for different combinations of features over Instagram data. Shaded regions in this table correspond to the best values of the four evaluation metrics considered in the experiment.

As can be seen from the Figure 5.3, the use of the complete set of features yields the highest accuracy and F-measure values and also the lowest FA rate for Instagram data. In fact, by using the eight input features, our approach achieved accuracy higher than 98% and F-measure over 0.98, both pointing at accurate results. The use of the eight user features yielded high CD rate (96.9%) and low FA rate (0.2%) suggesting their practical usability to accurately identify spammers.

On the other hand, we observed that, for some feature combinations, our approach was able to correctly identify all spammers (CD rate 100%) with the expense of also selecting 13.3% of legitimate users as spammers. In the combination of two features, the best F-measure (0.952) and accuracy (96.5%) achieved with the combination of idle time and proportion of bidirectional friends. We observed that, while more features are combined, better results are

Input features	Accuracy	CD	FA	F-measure
Proportion of bidirectional friend & Full name length	83.8%	69.0%	0.3%	0.815
Proportion of bidirectional friends & Total number of media posted	84.5%	69.9%	0.3%	0.821
Followers to following ratio & Idle time	90.1%	100%	13.3%	0.840
proportion of bidirectional friends & Idle time	96.5%	94.0%	1.9%	0.952
Proportion of bidirectional friends & Full name length & Idle time	97.1%	98.6%	3.5%	0.959
Average number of likes received per post & Average number of posts per week & Proportion of bidirectional friends & Full name length & Idle time	98.7%	99.5%	1.6%	0.982
Average number of likes received per post & Average number of posts per week & Proportion of bidirectional friends & Total number of media posted & Followers to following ratio & Full name length & Idle time	98.7%	99.5%	1.6%	0.982
All features	**98.7%**	**96.9%**	**0.2%**	**0.983**

Figure 5.3: Performance results over Instagram data.

achieved. For example, the combination of three features (full name length, idle time and proportion of bidirectional friends) yielded F-measure of 0.959, accuracy equals to 96.5%, and CD rate of 98.6%. As can be seen from the results, for two different feature combinations (selecting five and seven feature combinations) our approach reported the same CD rates (99.5%), FA rates (1.6%), accuracy (98.7%) and F-measure (0.982).

Overall, this experiment seemed to suggest that, in general, a substantial improvement is gained in identifying spammers by considering total number of media posted, average number of posts per week, average number of likes received per post, idle time, proportion of bidirectional friends, full name length, followers to following ratio, and account lifetime. In fact, as can be seen from Figure 5.3, the combination of these eight features yields the best trade-off between CD rate and FA rate to get higher F-measure and accuracy for Instagram data.

5.2.4 Experiment 2

The goal of this second set of experiments was to compare the performance of our approach with several machine learning algorithms. Note that, in this experiment we have used all the eight features as input of all competing algorithms. To begin with, we set $M_{max} = 4$ in all our experiments and selected the number of components that minimizes $ICL - BIC$. After running our proposed method on the Instagram data set, we observed that the users' feature vectors were well fitted by two distinctive Dirichlet components. As we expected the component which contained vectors with the lowest values corresponded to the class of spammers. Figure 5.4 illustrates the result of the compared algorithms. Shaded regions in this table correspond to the best values of accuracy, CD rate, FA rate and F-measure.

As can be seen from Figure 5.4, except RBF Network, all the remaining algorithms achieved accurate results. Our approach reached an accuracy of 98.7%, CD rate of 96.9%, FA rate of 0.2%, and F-measure of 0.982. As can be seen, the SVM reports the highest accuracy (99%), CD rate (98.3%), and F-measure (0.987). While our approach reports the lowest FA rate, F-measure and accuracy values which are very close to that of SVM. Figure 5.4 also suggests that meta classifiers as well as tree classifiers show acceptable performance that are fairly comparable to SVM and our approach. However, RBF network was not successful in accurately identifying spammers on Instagram. Overall, the experiment results in this part show that our proposed unsupervised method performs as well as (and sometimes better than) several supervised approaches.

5.3 Identifying Spammers on Twitter

In recent years, microblogging has become an important platform for social communications where people can seek and exchange real-time information. Twitter is the most popular microblogging service for sharing opinions, news, and trending topics that also presents a new level of communication in terms of social networks (Cheng et al., 2013). It is obvious that spamming activities explode once an online communication medium becomes popular, thus Twitter turned out to be a proper target for spammers to spread irrelevant and undesirable contents.

Algorithm	Accuracy	CD	FA	F-measure
Proposed	**98.7%**	**96.9%**	**0.2%**	**0.982**
AdaBoost	97.2%	96.9%	1.0%	0.976
Bagging	98.5%	97.4%	0.7%	0.980
Decorate	98.2%	96.5%	0.7%	0.976
LogitBoost	98.2%	96.5%	0.7%	0.976
MultiBoostAB	98.2%	96.1%	0.5%	0.976
ADTree	98.2%	97.4%	1.2%	0.976
Random Forest	98.1%	97.8%	1.7%	0.974
RBF Network	92.6%	83.8%	2.4%	0.891
SVM	99.0%	98.3%	0.5%	0.987

Figure 5.4: Accuracy of compared algorithms on Instagram data.

Users of Twitter post messages up to 140 characters, called tweets. These short messages consist of personal information about users, news or links to contents such as images, videos, articles. Users on Twitter are able to follow and/or being followed by other participants. They can also spread information by re-tweeting interesting tweets. A Twitter user can also be tagged by other users while their user-names preceded by the "@" symbol. All these features on Twitter bring out opportunities for spammers to take advantage of system such as promoting irrelevant businesses or posing as a normal user tweeting spam content periodically.

5.3.1 Twitter Data

Twitter seems well suited for the task of our study, as it contains a rich store of information and social interactions. To collect Twitter data, we used the search API of Twitter to gather tweets and users profiles. For the purpose of evaluation, two human annotators were recruited to analyze the collected data in order to produce a labeled collection of legitimate users and spammers. The labeling was done by looking at each user's profile and also by examining the top 20 recent tweets posted by each user. Note that the data collection and the labeling were performed in two phases. The first phase aimed to collect the likely legitimate users' profiles. To this end, we polled the Twitter Public Timeline through the search API of Twitter. User profiles were gathered one by one. The tweets and profile of these users were then manually

analyzed. As a result of this process, 526 users were labeled as legitimates. The second phase aimed at collecting potential spammers' profiles. To this end, we extracted tweets using keywords usually employed by spammers such as "win a free trip", "make money", "cheap viagra", "affiliate marketing", "mortgage", etc. Then, the collected tweets and the associated users' profiles were manually analyzed by the annotators. As a result of this process, 455 users were labeled as spammers. So, our hand-coded data sets contained 981 users; out of which 455 were labeled as spammers and the rest (526) as legitimate.

5.3.2 Twitter Social Behavior-based Features of users

Once the labeled collection of users was gathered, each user was presented by a feature vector which is composed of several attributes that reflect the user's legitimacy and reputation level on Twitter. It is important to note that, in this thesis we do not aim at defining new features to identify spammers on Twitter. This would be far beyond the scope of this thesis. In contrast to Instagram, the spam phenomenon in Twitter has been already investigated in previous works and appropriate features have been defined (Stringhini et al., 2010), (Lee et al., 2010). In our experiments, we thus utilize existing features that may characterize spammers. The main goal of these experiments is illustrating the suitability of our approach in handling spammers in another online social network which is different from Instagram. The set of features that we used is explained as follows:

- **Follower to following ratio**: Spammers tend to follow a large number of people while they are followed by a few participants, so this ratio is expected to be low for spammers.

- **Average time between tweets**: Spammers tend to post more tweets than normal users on average, over a period of a same time. This number is thus expected to be lower for spammers.

- **Number of mentions @ to the number of tweet ratio**: Legitimate users have more tendencies to use the mention @ than spammers, so this number is expected to be low for spammers.

- **Number of tweets to account lifetime ratio**: Spammers have a short account lifetimes but a large amount of tweets, so this ratio is expected to be higher for spammers.

Note that in this data set we assumed that spammers are characterized by features with small values. Hence, a linear inversion was applied to the feature with higher values, so that the inverted small values corresponded to the spammers.

5.3.3 Experiment 1

Similar to the experiments on Instagram data set, the aim of the first set of experiments was to evaluate our approach using different subsets of the Twitter features (follower to following ratio, average time between tweets, number of mentions @ to the number of tweet ratio, number of tweets to account lifetime ratio) considered in this thesis. In our experiments, we constructed several data sets using the following subsets of the features:

1. Number of tweets to account lifetime ratio and average time between tweets.
2. Number of mentions @ to the number of tweet ratio and number of tweets to account lifetime ratio.
3. Average time between tweets and follower to following ratio.
4. Number of mentions @ to the number of tweet ratio and follower to following ratio.
5. Number of mentions @ to the number of tweet ratio and follower to following ratio and average time between tweets.
6. Number of mentions @ to the number of tweet ratio and number of tweets to account lifetime ratio and average time between tweets.
7. The complete set of four features.

Then, for each constructed data set, we used our approach to identify spammers. To this end we set M_{max} to 4 in all our experiments and selected the optimal number of components that minimize $ICL - BIC$. Interestingly, as with the experiments on Instagram data set,

we found that the number of components varies from two to three. The Dirichlet component that represents the lowest feature values corresponds to spammers.

In Figure 5.5, we demonstrate the performance results over Twitter data, based on several features combinations. Shaded regions in the table correspond to the best values of the four evaluation metrics considered in this thesis. As can be seen from this table, the combination of the four Twitter users features provides the best accuracy (96.7%), CD rate (97.4%), and F-measure (0.975). On the other hand, the lowest FA rate achieves in combination of two features (number of mentions @ to the number of tweet ratio and number of tweets to account lifetime ratio), however the accuracy, F-measure and CD rate are relatively low. As can be seen from the results, the combination of two features (number of mentions @ to the number of tweet ratio and follower to following ratio) achieves fairly good results with accuracy equal to 92.7%, CD rate of 96.3% and F-measure of 0.896 in comparison with other combination of two features. In combination of three features, the FA rate decreased remarkably to 2.0% and even 1.7% but respectively F-measure, accuracy and CD rate have degraded. Overall, the first experiment suggests that the combination of the follower to following ratio, average time between tweets, number of mentions @ to the number of tweet ratio and number of tweets to account lifetime ratio provides the best trade-off between FA rate and achieves the highest accuracy, F-measure and CD rate.

5.3.4 Experiment 2

Our goal is now to compare the performance of our approach to that of AdaBoost, Bagging, Decorate, LogitBoost, MultiboostAB, ADTree, Random Forest, RBF Network and SVM. Note that, for all compared algorithms, we present results using the four Twitter users' features considered in this thesis (that is follower to following ratio, average time between tweets, Number of mentions @ to the number of tweet ratio and number of tweets to account lifetime ratio). Interestingly, as with the experiments on Instagram data, we observed that the user feature vectors are well fitted by two Dirichlet components. The component that contains vectors with the lowest values corresponds to spammers. Figure 5.6 illustrates the results of the compared algorithms. Shaded regions correspond to the best accuracy, CD rate, FA rates

Input features	Accuracy	CD	FA	F-measure
Number of tweets to account lifetime ratio & Average time between tweets	53.8%	3.7%	21.7%	0.500
Number of mentions @ to the number of tweet ratio & Number of tweets to account lifetime ratio	68.5%	53.2%	0.0%	0.694
Average time between tweets & Follower to following ratio	72.7%	60.7%	2.6%	0.749
Number of mentions @ to the number of tweet ratio & Follower to following ratio	92.7%	96.3%	9.0%	0.896
Number of mentions @ to the number of tweet ratio & Follower to following ratio and & Average time between tweets	72.7%	54.7%	2.0%	0.701
Number of mentions @ to the number of tweet ratio & Number of tweets to account lifetime ratio & Average time between tweets	91.3%	80.6%	1.7%	0.879
All features	**96.7%**	**97.4%**	**4.7%**	**0.975**

Figure 5.5: Performance results over Twitter data.

and F-measure values.

As can be seen from Figure 5.6, competing algorithms reported fairly accurate results. The best accuracy (98.4%) is achieved by ADTree and Random Forest. Random Forest also reports the highest CD rate (100%) and F-measure value (0.989). As depicted by Figure 5.6, such results are comparable to those achieved by our approach and also to those reported by meta classifiers and function-based classifiers. In fact, each algorithm reports accuracy greater than 96%, CD rate over 97%, and F-measure values more than 0.97. In terms of FA rate, the best result (3.7%) is reported by ADTree and the worst result (6.9%) is achieved by MultiBoostAB. Our approach reports a FA rate of 4.7% which is lower than the average FA rate of all competing algorithms, which is 5.13%.

To summarize, the experiments conducted on Instagram and Twitter data sets suggest that supervised algorithms as well as our unsupervised approach provide meaningful results. However, the advantage of our approach is that it performs spammer detection in an unsupervised fashion without relying on labeled data or any detection threshold required to be set by users. Supervised machine learning algorithms, however, suffer from their dependency on the training data which are more difficult and time consuming to obtain than unlabeled ones.

Algorithm	Accuracy	CD	FA	F-measure
Proposed	**96.7%**	**97.4%**	**4.7%**	**0.975**
AdaBoost	96.3%	98.2%	4.8%	0.973
Bagging	97.2%	98.7%	5.4%	0.979
Decorate	98.0%	99.2%	4.2%	0.986
LogitBoost	97.5%	98.7%	4.8%	0.982
MultiBoostAB	96.3%	97.9%	6.9%	0.973
ADTree	**98.4%**	99.5%	**3.7%**	0.988
Random Forest	**98.4%**	**100%**	4.8%	**0.989**
RBF Network	96.5%	97.4%	5.3%	0.874
SVM	96.1%	97.4%	6.3%	0.972

Figure 5.6: Accuracy of compared algorithms on Twitter data.

[Cette page a été laissée intentionnellement blanche]

CHAPTER VI

CONCLUSION

In this thesis, we have discussed some drawbacks of existing spammers detection approaches including their incapability to automatically discriminate between spammers and legitimate users, their dependency on labeled data, and their need for user threshold parameters which are difficult to tune. To address this problem, we have proposed a mixture model-based approach to automatically identify spammers in different online social networks. Our approach is a statistical framework based on Dirichlet mixture model which is able to automatically detect spammers without any prior knowledge or human intervention. We first proposed to represent each user as a feature vector such that each element of the vector contains information that would reflect the user's legitimacy and reputation level on the social media. Next, we modeled these vectors as a mixture of Dirichlet distribution. The number of component is estimated using the integrated classification likelihood Bayesian information criterion, while the parameters of the mixture are estimated using the EM algorithm. Such an approach allows the identification of the Dirichlet component containing the spammers. We evaluated the suitability of our approach in tests and comparisons with some supervised methods, using real data from Instagram and Twitter. The experiments showed that the proposed approach yielded high-quality results.

As a matter of fact, our unsupervised approach for spammers identification exhibits results that are comparable to those of supervised attribute learning algorithms. A general assumption about the supervised approaches would be that they performed better than unsupervised methods since they employed the grand truth provided by human in their procedure while unsupervised method do not have access to such information and should mine unlabeled data. However the experimental results regarding the efficacy of our approach show that our unsu-

pervised method performs as well as (and in several cases even superior) several supervised methods.

Finally, it is worth nothing that, in contrast to most existing spammers detection methods, our approach has several practical advantages. As discussed earlier, the proposed method is parameterless which is, in turn, a considerable advantage in practice. Parameter-laden methods are, however, critical and their application to real situations is not obvious since it is rarely possible for users to apply the parameters values accurately. Furthermore, the method presented in this thesis does not require labeled samples or prior knowledge about the data under investigation to detect spammers. In fact, our approach is able to automatically identify spammers from legitimate users in real-world scenarios, as it is shown on real data from Instagram and Twitter. We believe that these notable features of the proposed approach provide significant evidence about its practicality and should be considered to be a viable option in this regard.

BIBLIOGRAPHY

Abraham, A., Hassanien, A. E., and Snášel, V. (2009). *Computational Social Network Analysis: Trends, Tools and Research Advances*. Springer.

Adedoyin Olowe, M., Gaber, M. M., and Stahl, F. (2013). A survey of data mining techniques for social media analysis. *Computing Research Repository*, abs/1312.4617.

Aghaei, S., Nematbakhsh, M. A., and Khosravi Farsani, H. (2012). Evolution of the world wide web: From web 1.0 to web 4.0. *International Journal of Web & Semantic Technology*, 3(1).

Ahlqvist, T., Back, A., Halonen, M., and Heinonen, S. (2008). *Social media roadmaps: exploring the futures triggered by social media*. VTT.

Baruah, T. D. (2012). Effectiveness of social media as a tool of communication and its potential for technology enabled connections: A micro-level study. *International Journal of Scientific and Research Publications*, 2(5).

Bdiri, T. and Bouguila, N. (2012). Positive vectors clustering using inverted dirichlet finite mixture models. *Expert Systems with Applications*, 39(2):1869–1882.

Becker, H., Chen, F., Iter, D., Naaman, M., and Gravano, L. (2011). Automatic identification and presentation of twitter content for planned events. In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media*.

Benevenuto, F., G., M., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on twitter. In *Proceedings of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. (2009). Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 620–627.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Zhang, C., and Ross, K. (2008). Identifying video spammers in online social networks. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*, pages 45–52.

Bhat, S. Y. and Abulaish, M. (2013). Community-based features for identifying spammers in online social networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 100–107.

Bilge, L., Strufe, T., Balzarotti, D., and Kirda, E. (2009). All your contacts are belong to us: Automated identity theft attacks on social networks. In *Proceedings of the 18th International Conference on World Wide Web*, pages 551–560.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer Verlag New York Inc.

Bosma, M., Meij, E., and Weerkamp, W. (2012). A framework for unsupervised spam detection in social networking sites. In *Proceedings of the 34th European Conference on Information Retrieval*, pages 364–375.

Bouguessa, M. (2011). An unsupervised approach for identifying spammers in social networks. In *Proceedings of the 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, pages 832–840.

Bouguessa, M., Wang, S., and Sun, H. (2006). An objective approach to cluster validation. *Pattern Recognition Letters*, 27(13):1419–1430.

Bouguila, N., Ziou, D., and Vaillancourt, J. (2004). Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13(11):1533–1543.

Boutemedjet, S., Ziou, D., and Bouguila, N. (2010). Model based subspace clustering of non gaussian data. *Neuro computing*, 73(10-12):1730–1739.

Boyd, D. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Cheng, Z., Caverlee, J., and Lee, K. (2013). A content-driven framework for geolocating microblog users. *ACM Transactions on Intelligent Systems and Technology*, 4(1):2:1–2:27.

Figueiredo, M. A. T. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Graybill, F. A. (1983). *Matrices with applications in statistics*. Belmont, Calif.: Wadsworth International Group.

Han, F. and Liu, H. (2014). High dimensional semiparametric scale-invariant principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):2016–2032.

Hu, X., Tang, J., and Liu, H. (2014). Online social spammer detection. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 59–65.

Jin, X., Lin, C. X., Luo, J., and Han, J. (2011). Socialspamguard: A data mining-based spam detection system for social media networks. *Publication of the Very Large Database*, 4(12):1458–1461.

Kagdi, H., Collard, M. L., and Maletic, J. I. (2007). A survey and taxonomy of approaches for mining software repositories in the context of software evolution. *Journal of Software Maintenance and Evolution*, 19(2):77–131.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24.

Lee, K., Caverlee, J., and Webb, S. (2010). Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 435–442.

Ma, Z. and Leijon, A. (2009). Beta mixture models and the application to image classification. In *Proceedings of the 16th IEEE International Conference on Image Processing*, pages 2045–2048.

Ma, Z., Rana, K. P., Taghia, J., Flierl, M., and Leijon, A. (2014). Bayesian estimation of dirichlet mixture model with variational inference. *Pattern Recognition*, 47(9):3143–3157.

Markines, B., Cattuto, C., and Menczer, F. (2009). Social spam detection. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, pages 41–48.

Mislove, A. (2009). *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. PhD thesis, Rice University, Department of Computer Science.

Narisawa, K., Bannai, H., Hatano, K., and Takeda, M. (2007). Unsupervised spam detection based on string alienness measures. In *Proceedings of the 10th International Conference on Discovery Science*, pages 161–172.

Narisawa, K., Ikeda, D., Yamada, Y., and Takeda, M. (2006). Detecting blog spams using the vocabulary size of all substrings in their copies. In *In Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem*.

Nie, F., Yuan, J., and Huang, H. (2014). Optimal mean robust principal component analysis. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1062–1070.

O'Reilly, T. (2007). What is web 2.0: Design patterns and business models for the next generation of software. MPRA Paper 4578, University Library of Munich, Germany.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3):515–554.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348.

Pinheiro, C. A. R. (2011). *Social Network Analysis in Telecommunications*. Wiley.

Russell, S. J., Norvig, P., Candy, J. F., Malik, J. M., and Edwards, D. D. (1996). *Artificial Intelligence: A Modern Approach.* Prentice Hall Inc.

Shewmaker, J. (2014). Social media in the classroom: Challenges and strategies in faculty development. *The Journal of Social Media in Society*, 3(1).

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validatedlikelihood. *Statistics and Computing*, 10(1):63–72.

Soiraya, M. and Thanalerdmongkol, S. an Chantrapornchai, C. (2012). Using a data mining approach: Spam detection on facebook. *International Journal of Computer Applications*, 58(13):27–32.

Stringhini, G., Kruegel, C., and Vigna, G. (2010). Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 1–9.

Tan, E., Guo, L., Chen, S., Zhang, X., and Zhao, Y. (2012). Spammer behavior analysis and detection in user generated content on social networks. In *Proceedings of the 2012 IEEE 32nd International Conference on the Distributed Computing Systems*, pages 305–314.

Tan, E., Guo, L., Chen, S., Zhang, X., and Zhao, Y. (2013). Unik: unsupervised social network spam detection. In *Proceedings of the 22nd ACM International Conference on information & knowledge management*, pages 479–488.

Tseng, C. and Chen, M. (2009). Incremental svm model for spam detection on dynamic email social networks. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, pages 128–135.

Uemura, T., Ikeda, D., and Arimura, H. (2008). Unsupervised spam detection by document complexity estimation. In *Proceedings of the 11th International Conference on Discovery Science*, pages 319–331.

Viswanath, B., Bashir, M. A., Crovella, M., Guha, S., P., G. K., Krishnamurthy, B., and Mislove, A. (2014). Towards detecting anomalous user behavior in online social networks. In *Proceedings of the 23rd USENIX Security Symposium*, pages 223–238.

Wang, C. and Lee, M. K. O. (2014). Understanding microblog addiction on smartphone: the role of media characteristics. In *Proceedings of the International Conference on Mobile Business.*

Webb, S., Caverlee, J., and Pu, C. (2008). Social honeypots: Making friends with a spammer near you. In *Proceedings of the 5th Conference on Email and Anti-Spam.*

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. N., and Zhao, B. Y. (2009). User interactions in social networks and their implications. In *Proceedings of the 4th ACM European Conference on Computer Systems*, pages 205–218.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420.

Yoshida, K., Adachi, F., Washio, T., Motoda, H., Homma, T., Nakashima, A., Fujikawa, H., and Yamazaki, K. (2004). Density-based spam detector. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 486–493.

Zhao, H. V., Lin, W. S., and Liu, K. J. R. (2011). *Behavior Dynamics in Media-Sharing Social Networks.* Cambridge University Press.

Zhu, L., Sun, A., and Choi, B. (2011). Detecting spam blogs from blog search results. *Information Processing & Management*, 47(2):246–262.

Zhu, Y., Wang, X., Zhong, E., Liu, N., Li, H., and Yang, Q. (2012). Discovering spammers in social networks. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 171–177.